Are Timing-Based Side-Channel Attacks Feasible in Shared, Modern Computing Hardware?

Reza Montasari, School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
Amin Hosseinian-Far, Department of Business Systems and Operations, The University of Northampton, Northampton, UK
Richard Hill, Department of Computer Science, University of Huddersfield, Huddersfield, UK
Farshad Montaseri, Independent Researcher, Iran
Mak Sharma, School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
Shahid Shabbir, School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

ABSTRACT

This article describes how there exist various vulnerabilities in computing hardware that adversaries can exploit to mount attacks against the users of such hardware. Microarchitectural attacks, the result of these vulnerabilities, take advantage of microarchitectural performance of processor implementations, revealing hidden computing process. Leveraging microarchitectural resources, adversaries can potentially launch timing-based side-channel attacks in order to leak information via timing. In view of these security threats against computing hardware, the authors analyse current attacks that take advantage of microarchitectural elements in shared computing hardware. This analysis focuses only on timing-based side-channel attacks against the components of modern PC platforms - with references being made also to other platforms when relevant - as opposed to any other variations of side-channel attacks which have a broad application range. To this end, the authors analyse timing attacks performed against processor and cache components, again with references to other components when appropriate.

KEYWORDS

Attack Taxonomy, Hardware Vulnerabilities, Microarchitectural Attacks, Processor, Side-Channel Attacks

1. INTRODUCTION

Side-Channel Attacks, hereafter referred to as SCAs, pose serious security and privacy threats to modern and shared computing hardware (Ge et al., 2016; Liu et al., 2015; Xiao and Xiao, 2013; Kong, 2009). They are the result of spatial and temporal sharing of processor components between various applications as they run on the processor. A SCA – both theoretical (Hu, 1992, Page, 2002) and practical (Bernstein, 2005; Osvik et al., 2006) – is carried out through the exploitation
of inadvertent information leakage from computing hardware (Gruss, 2017; Spreitzer et al., 2016) or via the exploitation of Microarchitectural channels in order to deduce secret keys such as those utilised in symmetric cryptography (Inci et al., 2016; Yarom and Bengter, 2014; Zhang et al., 2014). For instance, through a SCA, an attacker will be able to exfiltrate secret keys used in cryptographic implementations, or gain information about it by probing the runtime. As an example, in 128-bit AES implementations that utilises four 1KB precomputed SBox tables such as OpenSSL (OpenSSL, 2016; Gullasch et al., 2011; Neve and Seifert, 2006), the probing of the ciphertext can result in the extraction of the complete secret key (Zhang et al., 2014; Agrawal and Mishra, 2012; Neve and Seifert, 2006). Various systems have inherent side-channel vulnerabilities that can be exploited by the attackers to launch devastating SCAs. For instance, an adversary can simplys carry out a differential power analysis (Kocher et al., 2017; Moradi et al., 2011; Kocher et al., 2011; Barendghi et al., 2010; Coppens et al., 2009; Schramm et al., 2004; Guilley et al., 2004; Kocher et al., 2004) or monitor electromagnetic radiation (Longo et al., 2015; Hayashi et al., 2013; Homma et al., 2010), etc., in order to deduce vital data from the victims’ systems (Zhang and Lee, 2014).

Furthermore, processor architecture features including simultaneous multithreading (Tromer et al., 2010; Aciicmez et al., 2007; Percival, 2005), control speculation and shared caches (Steffan et al., 2000; Tsai and Yew, 1996) can unintentionally accelerate side channels or enable new side channels (Yarom and Falkner, 2014; Wang and Lee, 2006). As a result, attackers can detect and exploit contention between hardware threads on the multiplier unit (Ge et al., 2016; Ghan et al., 2015; Chen and Venkataramani, 2014). Such contention can be also exploited to create a side channel (Liu et al., 2016; Hunger et al., 2015; Ristenpart et al., 2009), for instance, to enable a malicious thread to differentiate multiplications from squaring in OpenSSL’s RSA implementation (Aciicmez et al., 2007; Wang and Lee, 2006). These attacks can determine the latency which result from contentious threats that are made to wait for access to functional units (Ge et al., 2016; Tromer et al., 2010; Ristenpart et al., 2009).

SCAs have increasingly advanced from attacks on computing devices to attacks on general-purpose computing platforms (Spreitzer et al., 2016) and cloud computing infrastructures (Liu et al., 2015; Zafirt, 2015; Zhang et al., 2014), and finally to attacks on mobile platforms (Lipp et al., 2016; Song et al., 2016; Chen et al., 2014; Sarwar et al., 2013). Such advancement coupled with evolution in computational power of modern processors (as a result of the sharing of processor units) have provided researchers with new research opportunities in the new field of Microarchitectural analysis to continue to devise new and more sophisticated SCAs (such as that of exploiting the sharing features of processors) as well as Covert Channel Attacks, hereon abbreviated to CCAs. For instance, studies since as far back as 1973 have demonstrated that microprocessors and caches are susceptible to both SCAs and CCAs (Gruss et al., 2017; Pessl et al., 2016; Liu et al., 2015; Hund et al., 2013; Kim et al., 2012; Neve et al., 2006; Osvik et al., 2006; Bernstein, 2005; Percival, 2005; Tsunoo et al., 2003; Kelsey et al., 2000; Kocher, 1996; Hu, 1992; Wray, 1992; Lampson, 1973) by illustrating that cache architecture brings about inconsistency in the runtime because of dissimilar memory accesses. Such vulnerability poses threats to hardware because cache accesses are reliant, for instance, on the inputs of plaintext and the key (Crane et al., 2015; Tromer et al., 2010; Neve and Seifert, 2006).

Furthermore, as well as modern computing processors and caches, SCAs can also be carried out against public key cryptography schemes (Zhang et al., 2014; Bellare et al., 2013), AES (Irazoqui et al., 2015, a; Irazoqui et al., 2014), ECDSA (Benger et al., 2014), TLS messages (Irazoqui et al., 2015, b), items in a shopping cart (Zhang et al., 2014) or even the key strokes typed in a keyboard (Gruss et al., 2015). In addition, SCAs can also be mounted against PaaS clouds (Liu et al., 2015; Zafirt, 2015; Zhang et al., 2014), across processors (Irazoqui et al., 2016; Crane et al., 2015; Hund et al., 2013) and in smartphones (Lipp et al., 2016; Song et al., 2016; Chen et al., 2014). Having examined the state-of-the-art reveals that on one hand processor manufacturers incorporate new enhanced security features, but on the other hand security researchers compromise these features just to demonstrate the way in which secret information can be emitted (Grus et al., 2016; Hunger et al., 2015; Wang et
al., 2014; Saltaformaggio et al., 2013; Acicmez and Seifert, 2007; Wang and Lee, 2006). In light of such increasing attention to both SCAs and CCAs, researchers have proceeded to suggest various hardware and software countermeasures to deal with these attacks (Martin et al., 2012; Kong et al., 2009; Wang and Lee, 2007; Page, 2005).

1.1. Key Contributions and the Methodology Followed

In light of the above discussion, this study thoroughly reviews and analyses both sides of the competition, i.e. attacks and countermeasures, and makes the following contributions that include:

1. A systematic survey of existing literature within the context;
2. Analysing the field of Microarchitectural Analysis;
3. Describing the main characteristics of Microarchitectural elements that enable the exposure of side channels;
4. Identifying and examining existing Microarchitectural SCAs;
5. Summarising the main features of hardware that bring about side channels, in order to enable the research communities and software developers to gain better insight into the way that applications can be pro-actively designed to prevent SCA vulnerabilities.

Therefore, as already stated, this study examines SCAs with the main focus being on Timing-Based Side-Channel Attacks, hereon referred to as TBSCAs, and identify common features between them. The study then proceeds to analyse existing countermeasures and propose new ones. From our analysis, we deduce insight in relation to the current state of knowledge observed in the literature, establish means of attacks, predict possible future modus operandi of attacks, and propose effective future directions for the development of appropriate defence mechanisms. Although we present theoretical work of clear relevance, we primarily focus our attention on practical and established attacks and defence mechanisms. We believe that the understanding obtained from this study enables researchers to establish directions for future research and also to address such attacks on a larger scale.

1.2. Scope of the Survey

Based on our survey of the literature, various Microarchitectural SCAs have been identified, which organise into a taxonomy of 13 general sub-categories, including: Acoustic Cryptanalysis Attack, Branch-Prediction Attack, Cold Boot Attack, Cache Attack, Differential Fault Analysis Attack, DMA Attack, Electromagnetic Attack, Fault-Attacks, Lucky-Thirteen Attack, Pass the Hash Attack, Power-Analysis Attack, Tempest Attack, and Timing Attack. The emphasis of this study is only on Timing-Based Side-Channel Attacks, as opposed to any other 12 variants, with brief reference to other variations only when appropriate. Furthermore, the emphasis of our study on TBSCAs is only on those Timing Attacks that are capable of compromising ‘the components of a PC platform’ (such as a hard drive or a modern processor that can consist of processor cores and any functional units inside a multi-core multi-threaded processor) and ‘entities in a network’. Again, references are made also to TBSCAs in other platforms such as mobile devices or cloud infrastructure only when appropriate. In addition, this study does not explore covert channels even though they are referred to when necessary or appropriate. Any other topics related to side channels are beyond the scope of this paper. Existing countermeasures against TBSCAs within PC platforms and entities in networks are then analysed, and new strategies are proposed.

1.3. Outline of the Paper

The remainder of the paper is structured as follows: Section 2 provides a background for Microarchitectural Analysis. In Section 3, Timing-Based Side-Channel Attack are presented and examined in detail, while in Section 4, Side-Channel Attacks against RSA implementations are
analysed. Finally, Section 5 concludes the study by providing a detailed discussion about trends in attacks, challenges to offset them and the future research direction in this research field. Two main contributions of this paper are the scope of the discussion, since few works of similar scope currently exist, and the provision of an agenda for the direction of future research.

2. BACKGROUND TO MICROARCHITECTURAL SIDE-CHANNEL ATTACKS

In 1996, Kocher (1996) demonstrated that attackers could potentially deduce RSA keys, named after RSA’s creators “Rivest-Shamir-Adleman” (Rivest et al., 1978), and also other decipher cryptosystems by carefully calculating the amount of time needed to conduct private key operations. Kocher was able to define SCAs as a method that enables adversaries to extract secret information utilised in a computing process from unintended impacts that the computing process has on its environment (Gruss, 2017; Crane et al., 2015; Genkin et al., 2014). His seminal work can be considered as a foundation for a whole new domain of research into Side-Channels (Gruss, 2017). Kocher carried out an attack which researchers now define as Timing-Based Side-Channel Attacks, attacks that take advantage of differences in runtime of a computing process. Kocher illustrated that a Side-Channel Attack against a weak system would not be computationally difficult and often necessitate only ciphertext. His study revealed that attackers could make relatively precise timing calculations that would result in breaking systems such as “cryptographic tokens, network-based cryptosystems”, and other applications (Kocher, 1996). Other seminal works in the field of SCAs include those by Mangard et al. (2008); Quisquater and Samyde (2001), Chari et al. (1999) as well as another study by Kocher himself in Kocher et al. (1999).

In the years that have followed, researchers have been able to illustrate SCAs based on changes in different computing settings (Gruss et al., 2017; Spreitzer and Plos, 2013), including: AES (Zhang et al., 2016, b; Spreitzer and Plos, 2013; Gullasch et al., 2011; Osvik et al., 2006), differential power analysis (Kocher et al., 2017; Kocher et al., 2011; Barenghi et al., 2010; Guilley et al., 2004; Schramm et al., 2004; Kocher et al., 2004), monitor electromagnetic radiation (Longo et al., 2015; Hayashi et al., 2013; Homma et al., 2010), sound and electromagnetic emission (Faruque et al., 2016; Genkin et al., 2014; Callan et al., 2014; Cai and Chen, 2011), photonic side-channel leakage emission (Carmon et al., 2017; Krämer et al., 2013; Schlösser et al., 2012) and many more. All such attacks necessitate adversaries to be able to have a physical access to the target device so as to monitor and deduce the secret information (Gruss, 2017; Ge et al., 2016; Spreitzer et al., 2016; Liu et al., 2015). However, the latest and more advanced SCAs including Cache-Timing Attacks (Ge et al., 2016; Yarom and Falkner, 2014; and Tromer et al., 2010) and DRAM row buffer attacks (Gruss, 2017; Schwarz et al., 2017; Pessl et al., 2016) can be carried out remotely by running malicious software within a cloud setting (Spreitzer et al., 2016; Xiao et al., 2016; Irazoqui et al., 2016). With the emergence of cloud computing phenomenon, the extent of SCAs has also evolved considerably since 2000s (Spreitzer et al., 2016; Kim et al., 2012). Likewise, with the rapid advancements in mobile technology, researchers have been able to demonstrate even more sophisticated SCAs compromising smartphones (Spreitzer et al., 2016; Song et al., 2016; Sarwar et al., 2013; Owusu et al., 2012; Lange et al., 2011). For instance, new attacks (Simon et al., 2016; Aviv et al., 2012; Xu et al., 2012; Cai and Chen, 2011) enable adversaries to deduce keyboard input on touchscreens through “sensor readings from native apps” (Spreitzer et al., 2016; Kambourakis et al., 2016; Aviv et al., 2012). Because typing on various places on the screen creates different vibrations, data from Motion (Cai and Chen, 2011), a SCA on touch screen smartphones with soft keyboards data, can be employed by an attacker to deduce the keys being typed. One of the methods to deduce keystrokes via the Motion, is to utilise a mobile application such as TouchLogger, an Android application that derives “features from device orientation data” (Cai and Chen, 2011). More advanced and new attacks can also enable the attackers to infer a user’s geographical location through the power consumption (Spreitzer et al., 2016; Mangard et al., 2008) and a victim’s identity through the proofs (Spreitzer et
al., 2016; Zhou et al., 2013) that is available from the proc filesystem (procfs) (Spreitzer et al., 2016; Michalevsky et al., 2015).

In the following section, we shall analyse various TBSCAs in relation to components of modern and shared PC platforms with references made also to other platforms (such as cloud computing and smart mobile phones) when relevant. To this end, a particular focus will be placed on TBSCAs with again making references to other variations of SCAs only when appropriate.

### 3. TIMING-BASED SIDE-CHANNEL ATTACKS

In this section, following giving a brief overview of timing channels and TBSCAs, we analyse various TBSCAs and demonstrate that hardware vulnerabilities resulting from various factors such as optimisations on a microarchitectural layer can be exploited by the attackers to launch devastating TBSCAs and as a result compromise the system security.

#### 3.1. Overview

Over the past few years, information leakage via covert channels and side channels have been a grave concern for security researchers (Wu et al., 2015; Luo et al., 2011; Chen et al., 2010). This issue has recently been exacerbated by certain features of modern processor architecture (Liu et al., 2015; Yarom and Falkner, 2014; Hund et al., 2013; Wang and Lee, 2006) such as simultaneous multithreading (Zhang and Reiter, 2013; Domnitser et al., 2012; Tromer et al., 2010), speculative memory accesses (Doychev et al., 2015; Guan et al., 2015; Yarom and Falkner, 2014; Harnik et al., 2010) and shared caches (Gruss et al., 2016; Crane et al., 2015; Zhang and Lee, 2014; Zhang et al., 2012) that can accelerate both covert channels and side channels.

Timing channels that represent both covert and side channels are the focus of this study as stated previously. A timing channel is a communication channel that can transfer information to a recipient and decoder by controlling the timing performance of an object (Biswas et al., 2017; Rowland, 1997) such as inter-packet delays of a packet stream (Wu et al., 2015; Liu et al., 2010; Sultana et al., 2013) or the reordering packets in a packet stream (Biswas et al., 2017; Sultana et al., 2013; Lu et al., 2010; Tsai et al., 2010). Such channels can be considered as a type of computer security attack that enables an adversary to develop an ability to transfer information objects between processes that are not intended to be allowed to communicate by the computer security policy.

The term “timing channel” was coined in 1973 by Lampson (Lampson, 1973) as channels that “are not intended for information transfer at all, such as the service program’s effect on system load” to differentiate it from benign channels that are exposed to access controls by computer security. Girling (1987) was first to investigate the usage of delays between packets transferred over computer networks for covert communication. This seminal study became the foundation for many other studies (Wendzel et al., 2015; Mazurczyk et al., 2014; Geddes et al., 2013; Luo et al., 2008; Zander et al., 2007; Partan et al., 2007; Elson et al., 2002; Ahsan, 2002) to identify and analyse timing channels. There exist three different widely-recognised types of Timing Channels, including: Covert Communications (Chen and Venkataramani, 2014; Gianvecchio and Wang, 2011; Liu et al., 2009), Timing-Based Side-Channels (Liu et al., 2015; Meyer et al., 2014; Hund et al., 2013; Stefan et al., 2013), and Network Flow Watermarking (Biswas et al., 2017, Bates et al., 2012; Zander et al., 2007).

In this study, we only focus on providing a detailed analysis of TBSCAs that we classify into two general categories. These include: (1) Timing-Based Side-Channel in a network, where an active entity within a network system communicates with other objects in the network, and in-system Timing-Based Side-Channel in a PC platform, where entities communicate with each other within the PC platform.

A TBSCA represents a type of SCA that exploits differences in the runtime of an algorithm (Brunley and Tuveri, 2011; Acićmez et al., 2005; Kocher, 1996). This denotes that by taking advantage of such differences, an adversary can potentially compromise a cryptosystem through the observation of the time required to run cryptographic algorithms (Pornin, 2017; Schneier,
2005; Kocher, 1996). Unlike cryptanalysis which is focused on the mathematics (Song et al., 2013; Otmani et al., 2010), for instance, in differential and linear cryptanalysis (Bogdanov and Rijmen, 2014; Mouha et al., 2011), a TBSCA is based on implementation and applies additional information collected from attacking such implementations (Snow et al., 2013; Hund et al., 2013; Sarwar et al., 2013; Kocher et al., 2011). Therefore, a TBSCA can also exploit the data-dependent performance features of the execution of an algorithm (Pornin, 2017; Chen et al., 2013; Kocher, 1996) as opposed to the mathematical components of the algorithm itself.

Furthermore, contrary to Cache-Based Side-Channel Attacks (CBSCAs), which exploit operational aspects of a system (Such as general-purpose systems) (Ge et al., 2016; Crane et al., 2015; Wang and Lee, 2007), TBSCAs are carried out via timing variation, even when the execution performance of the system is entirely known, and even when there is formal proof of the lack of Cache Channels (Ge et al., 2016; Murray et al., 2013; Schaefer et al., 1977). The time that each logical execution takes in a computer system to run varies according to the input (Patterson and Hennessy, 2017; Kirsch and Sokolova, 2012). With the exact analysis of the time for each execution, an adversary will be able to work backwards to the input (Vétillard and Ferrari, 2010; Kocher, 1996).

Calculating the time that a computer system takes to address specific queries can cause an emission of information from the system (Seibert et al., 2014; Weiß et al., 2012; Tromer et al., 2010; Hopper et al., 2010). The degree to which this information can assist an adversary will be based on certain factors such as crypto system implementation, the algorithms utilised, the CPU running the system, various execution details, timing attack remedies, the precision of the timing measurements, etc.

In addition, TBSCAs can exploit the effects of variations in encryption time caused by conditional branches that occur during encryption processing (Lee et al., 2016; Lawson, 2009; Aciqmez et al., 2007; Zhou and Feng, 2005; Tsunoo et al., 2003; Kocher, 1996). CPU cache (between the CPU and main memory) misses are also capable of creating such variations (Braun et al., 2015; Bonneau and Mironov, 2006). If the CPU accesses data that does not reside in the cache, a delay will be triggered because the target data must be loaded from main memory into the cache (Irazioqui et al., 2016; Gruss et al., 2016; Tsunoo et al., 2003). Therefore, the measurement of such delay can allow adversaries to establish the occurrence and frequency of cache misses. Last, but not least, Timing Attacks can also be facilitated through shared memory controllers (Ge et al., 2016; Pessl et al., 2016; Wang et al., 2014), where the computing hardware emits portions of its internal state such as confidential information via variations in performance and timing (Shafiee et al., 2015; Wang et al., 2014).

### 3.2. Prime+Probe Attack

In this attack, the adversary fills a cache set with his own lines, then waits for a specific period and proceeds to establish whether the lines are still cached (Gruss et al., 2017; Crane et al., 2015; Irazioqui et al., 2015, a; Tromer et al., 2010). The adversary will then be able to determine whether the victim accessed the designated cache set in the meantime (Gruss et al., 2017; Inci et al., 2016; Irazioqui et al., 2015, c; Apecechea et al., 2014). This requires that the attacker examines certain cache sets to establish the presence of a cache miss. To do so, he will need to time the accesses to the cache set after the victim executes (Inci et al., 2016; Crane et al., 2015). He assigns a group of cacheline-sized, cacheline-aligned memory blocks in order for such memory blocks to fill a collection of targeted cache sets (Zhang et al., 2016, a; Yarom and Falkner, 2014; Percival, 2005). Having done this, the attacker will then constantly carry out two attack phases including ‘prime phase’ and ‘probe phase’ (Gruss et al., 2017; Crane et al., 2015; Tromer et al., 2010). Within the ‘prime phase’, the attacker examines each memory chunk to eject all the victim’s data in such cache arrays. They will then wait for a delay time prior to carrying out the ‘probe phase’, where each memory chunk is examined in the group again and the time of memory accesses is determined (Gruss et al., 2017; Crane et al., 2015; Liu et al., 2015; Varadarajan et al., 2014). Longer access times represent one or more cache misses. This means that this cache array has been accessed by the victim between the prime and probe phases. The attacker will repeat these two phases many times in order to acquire traces that might overlap
with the victim’s performance of cryptographic operations (Zhang et al., 2016, a; Liu et al., 2015; Wei et al., 2012; Zhang et al., 2012, a; Ristenpart et al., 2009). As a result, the adversary will be able to establish which cache lines were replaced by the victim and infer more details concerning which addresses the victim accessed. To determine the cache lines that were replaced, the attacker will need to measure the speed of each cache access.

It should be noted that the Prime+Probe Attacks can target both the L1 through the use of cache sets and the Addressing Scheme (Oren et al., 2015; Yarom and Benger, 2014; Apeccechea et al., 2014; Yarom and Falkner, 2014; Zhang et al., 2012; Tromer et al., 2010; Aciciçmez, 2007; Osvik et al., 2006; Percival, 2005), as well as the Last Level Cache (LLC) (Irazoqui et al., 2015, a; Liu et al., 2015; Oren et al., 2015; Benger et al., 2014; Ristenpart et al., 2009). To perform the attack on L1, both the attacker and the victim will need to have access to the same physical CPU core concurrently. In contrast, to carry out the Prime+Probe Attack against L3, which is a more advanced version, both the adversary and the victim must be using the same CPU but not “necessarily” the CPU core (Inci et al., 2016; Irazoqui et al., 2014; Yarom and Falkner, 2014; Benger et al., 2014). Many of the advanced Prime+Probe Attacks do not require to be dependent upon De-Duplication (Liu et al., 2015, Irazoqui et al., 2015, a) or core sharing, resulting in them becoming broadly relevant (Inci et al., 2016).

Adversaries can also attack AES in OpenSSL 0.9.8 with Prime+Probe on the L1 data-cache (D-cache) (Liu et al., 2016; Kong et al., 2013; Wei et al., 2012; Brumley, 2011; Tromer et al., 2010; Osvik et al., 2006) and L1 instruction-cache (I-cache) contention (Irazoqui et al., 2015, a; Liu et al., 2015; Aciciçmez, 2007) to establish an end-user’s control row (Ge et al., 2016; Jia et al., 2014). This will enable the attacker to differentiate squares and multiples in OpenSSL 0.9.8d RSA (Allan et al., 2016; Liu et al., 2016).

Moreover, an adversary can launch a Prime+Probe Attack against elliptic-curve cryptography (ECC) in OpenSSL 0.9.8k using channel measurements (Garcia and Brumley, 2016; Allan et al., 2016; Ge et al., 2016; Yarom and Falkner, 2014; Yarom and Benger, 2014; Brumley and Hakala, 2009). CPU caches are a potent source of information leakage. Therefore, Prime+Probe Attacks can be mounted against them through manual identification of susceptibilities such as data accesses or instruction execution (Benger et al., 2014; Chen et al., 2013; Gullasch et al., 2011; Brumley and Hakala, 2009; Bonneau and Mironov, 2006; Osvik et al., 2006). In addition, Prime+Probe Attack can also be carried out within cloud computing environments. The cross-VM leakage exists in public clouds and is often a practical attack vector for stealing sensitive data (Inci et al., 2016; Green, 2013; Zhang et al., 2012, a). In the cross-VM context, the adversary and victim have two distinct VMs running as co-tenants on the same server. Thus, the adversary will be able to acquire co-tenancy of a malign VM on the same server as a target (Varadarajan et al., 2014; Xiao and Xiao, 2013; Zhang et al., 2012, a; Ristenpart et al., 2009). For instance, using a Prime+Probe technique, a cross-VM attack can be carried out to obtain ElGamal secret keys from the victim (Irazoqui et al., 2016; Osvik et al., 2006).

### 3.3. Time Slicing Attack

By performing a Time Slicing Attack (TSA), an adversary will be able to extract kernel and user-level ASLR offset on the branch target buffer (BTB) (Ge et al., 2016; Evtushkin et al., 2016; Hund et al., 2013; Aciciçmez et al., 2007, Hu, 1992). An Address Space Layout Randomisation (ASLR), first designed and coined by Linux PaX project (PaX, 2001), is a security technique used to prevent exploitation of memory vulnerability in operating systems that guard against buffer-overflow attacks. To provide such security, ASLR functions by randomising the location in which system executables are loaded into memory (Symantec, 2017; Davi et al., 2015; Shacham et al., 2004) and the offset of key program segments in virtual memory (Crane et al., 2015; Backes and Nürnberger, 2014; Bhatkar et al., 2003). In theory, this should render it difficult for the attacker to deduce addresses of certain code objects (Gruss et al., 2016; Evtushkin et al., 2016). However, as stated above, attackers can undermine the ASLR by mounting a TSA. For instance, a local attacker with restricted privileges can exploit the limitations of kernel space ASLR to launch a TSA against the memory management
system in order to infer information about the privileged address space layout (Hund et al., 2013; Bhatkar et al., 2003). Furthermore, through a TSA combined with another variant of TBSCA (See sub-section 3.7), namely Flush+Reload (Yarom and Falkner, 2014), the adversary will be able to exploit the Translation Lookaside Buffer (TLB) (Wang et al., 2017; Ge et al., 2016; Grunwald and Ghiasi, 2002) contention to overcome ASLR (Jang et al., 2016; Ge et al., 2016; Liu et al., 2015). This can result in the detection of both kernel-level and user-level virtual address space layout on ASLR-enabled Linux platforms since such platforms employ only portion of the virtual address bit as hash tags (Evyushkin et al., 2016; Wang et al., 2015; Larsen et al., 2014). In addition, invalid mappings that are not loaded into the TLB can also be exploited by the attacker, resulting in a void address which will activate another table walk (Ge et al., 2016; Evyushkin et al., 2016; Hund et al., 2013).

3.4. Remote Timing-Based Side-Channel Attacks

A Remote Timing-Based Side Channel Attack (RTBSCA), which is carried out within a network setting, enables an attacker to exploit weaknesses of a cryptographic design remotely (Hunger et al., 2015; Liu et al., 2015; Aciiçmez et al. 2007; Brumley and Boneh, 2005). A RTBSCA often remains undetected for a long time before its presence can be detected and the emitted private information can be decoded (Biswas et al., 2017; Lawson, 2009). Usually, the observer will not be able to enhance such properties since the private information is emitted by a defective operation (Biswas et al., 2017; Hunger et al., 2015). Achieving higher Timing-Based Side-Channel capacity is difficult since many constant observations are needed to decrease the error probability to enhance efficiency (Liu et al., 2015; Wu et al., 2015; Kocher, 1996). Acquiring high bandwidth necessitates optimising synchronization (Wu et al., 2015; Liu et al., 2015; Karlof and Wagner, 2003; Katabi, 2003). This denotes that matching clocks in the sender and receiver in order for them to correspond on the time duration for each bit (Hunger et al., 2015; Rhee et al., 2009). Synchronisation allows the sender and receiver to employ basic binary signalling without requiring self-clocking codes and yet obtain low bit error rates (Maurice et al., 2017; Hunger et al., 2015; Welzl, 2012; Welzl, 2005).

Often, there are vulnerabilities associated with SSH that an attacker can exploit to exfiltrate passwords remotely during an SSH session (Biswas et al., 2017; Balduzzi et al., 2012; Song et al., 2001) and private keys from an OpenSSL-based web server (Liu et al., 2015; Brumley and Boneh, 2005). For instance, these include: (1) the exposure of the original data size by the 8-byte limit of transferred packets (Seibert et al., 2014; Aciiçmez et al., 2010; Song et al., 2001) and (2) the emission of the inter-keystroke timing information since each portion of keystroke information is transmitted to the remote machine while in interaction state (Biswas et al., 2017; Seibert et al., 2014; Aciiçmez et al., 2011; Foo Kune and Kim, 2010; Raymond, 2001). Furthermore, under this attack, the adversary can exfiltrate a considerable amount of information that the victim types by employing sophisticated statistical methods (Chen et al., 2010). The attacker will be able to deduce secret information and values from the inter-keystroke timings by utilising, for instance, Hidden Markov Analysis Model (a technique employed to predict the value of a variable, the future value of which remains independent of its past history) and their key forecast algorithm (Biswas et al., 2017; Seibert et al., 2014; Aciiçmez et al., 2010; Foo Kune and Kim, 2010; Raymond, 2001).

Furthermore, adversaries can also launch a remote TBSCA against OpenSSL by exploiting the inherent susceptibility that exists in OpenSSL (in the Montgomery ladder in the Elliptic Curve Cryptosystem) in order to extract the secret key of a TLS server (Benger et al., 2014; Yarom and Benger, 2014; Brumley and Tuveri, 2011). Network Tomography, an essential part of network measurement, is responsible for performing traffic analysis by observing the network to ensure that all the links in a network are healthy (Mardani and Giannakis, 2016; Chawla et al., 2012; Danezis and Clayton, 2007). This is performed through the use of end-to-end queries that are transmitted by agents residing at vantage points in the network (Gong et al., 2012; Shmatikov and Wang, 2006). Using this same approach, that necessitates direct monitoring of network connections at local vantage points, an attacker will be to perform network analysis and as a result to launch a devastating TBSCA against
the hardware agents in the network (Ge et al., 2016; Meyer et al., 2014; Brumley and Boneh, 2005). In unusual cases in which the attacker is ‘extremely savvy’, he can also employ the same convention to launch a remote TBSCA by exploiting a scheduler (which is a side channel) between himself and the victim (Varadarajan et al., 2014; Wang et al., 2014; Stefan et al., 2013; Gullasch et al., 2011). Our point is substantiated by a study (Gong and Kiyavash, 2013) in which researchers were able to demonstrate that the attacker could establish the entire pattern of the victim by using “Shannon equivocation as a privacy metric” (Gong and Kiyavash, 2013). This attack is made possible if the scheduler is based on the policy of a first-come, first-served basis (Wang et al., 2014; Kadloor et al., 2010; Gupta, 2007). Again, using the same approach, the savvy attacker can launch a remote TBSCA by taking advantage of the Timing Side Channels inside a “home digital subscriber line (DSL) router” (Gong and Kiyavash, 2013; Kurose and Ross, 2010). Performing this attack, the adversary will be able to acquire the victim’s secret data such as passwords and voice over IP (VoIP) conversations (Biswaes et al., 2017; Lee et al., 2015; Mohaban et al., 2007).

3.5. Access-Driven Cache-Timing Attack

An Access-Driven Cache-Timing Attack (ADCTA) is another variation of SCAs that takes advantage of the emission of the memory locations that the victim process accesses (Kim et al., 2012; Gullasch et al., 2011). This attack involves probing the cache’s timings as a source of information emission (Zhang et al., 2012; Neve and Seifert, 2006). In the ADCTA, the cache performance is inspected with a fine granularity (Crane et al., 2015) as opposed to assessing the overall runtime. Using an ADCTA, an adversary will be able to determine whether a cache line has been ejected or not (Crane et al., 2015; Irazoqui et al., 2015; Acićmez and Koç, 2006). Furthermore, a savvy attacker can also potentially perform an ADCTA on the Advanced Encryption Standard (AES) block cipher (Irazoqui et al., 2015; Neve and Seifert, 2006; Bonneau and Mironov, 2006) by utilising compressed tables (Zhang et al., 2012; Gullasch et al., 2011) to extract the complete secret key in real time (for instance, for AES-128) (Crane et al., 2015). To do so, he will need only a restricted number of observed encryptions (without requiring any information concerning plaintext) to be able to exfiltrate the entire key due to, for instance, round analysis from the ciphertext.

ADCTAs are also made possible in cloud computing environments due to the vulnerability in certain hardware components such as the scheduler of the Xen hypervisor (Yarom and Falkner, 2014; Barham et al., 2003). For instance, the attacker can utilise a malicious virtual machine that will enable him to extract detailed, precise information from a victim VM that is running in parallel on the same computer (Yarom and Falkner, 2014; Zhang et al., 2012; Kim et al., 2012). As an example, to perform an ADCTA on a symmetric multiprocessing system (Braun et al., 2015; Winder, 2012), the adversary will require to deal with various challenges such as core migration (Winder, 2012; Bertozzi et al., 2006), multiple sources of channel noise (Braun et al., 2015; Winder, 2012) and also the problems with pre-empting the victim with adequate frequency to acquire detailed information from it (Zhang et al., 2012; Winder, 2012; Bertozzi et al., 2006). However, the attacker can bypass such challenges by utilising, for instance, libgcrypt cryptographic library to exfiltrate an ElGamal (ElGamal, 1985) decryption key of a GnuPG decryption (Yarom and Falkner, 2014), which is running in another guest, from the victim. The ADCTA can also be performed against certain OpenSSL (e.g. 0.9.8n) implementation of AES on Linux systems (Crane et al., 2015; Liu et al., 2015; Brumley and Boneh, 2005). In addition, it can also be used to mount a denial of service (DoS) attack on the task scheduler of Linux systems that allows the attackers to monitor all memory accesses of a victim process (Zhang et al., 2012; Gullasch et al., 2011).

Likewise, adversaries might be able to perform the ADCTA on a time-shared core to take advantage of a shared LLC. In such a case, they will need to exploit the cupid instructions or leverage fence instructions to be able to synchronise the instruction stream (Yarom and Falkner, 2014; Gullasch et al., 2011). Similarly, a successful ADCTA can break the isolation feature of system virtualisation (Yarom and Falkner, 2014; Kim et al., 2012; Zhang et al., 2012). In this situation, by employing and
adapting Bernstein’s attack’s link (Bernstein, 2005) in CTA (Weiβ et al., 2012), the attacker will be able to derive secret information from an isolated execution area. An ADCTA can also be performed to exploit the hardware-assisted multi-threading (Osvik et al., 2006; Percival, 2005) or single-threading (Neve and Seifert, 2006) ability of certain microprocessors so as to execute a spy process quasi in parallel to a cryptography process.

3.6. The Flush+Reload Technique

The Flush+Reload Attack (Yarom and Falkner, 2014), a variation of Prime +Probe Attacks (Irazoqui et al., 2015; Tromer et al., 2010), is based on the sharing of pages between the malicious and victim processes (Gruss et al., 2015; Liu et al., 2015; Yarom and Falkner, 2014; Osvik et al., 2006). By performing this attack, an adversary will be able to eject a particular memory line from the entire cache hierarchy through shared pages (Irazoqui et al., 2015; Yarom and Falkner, 2014). The Flush+Reload Attack has been adapted from Gullasch et al.’s (2011) technique for usage in both virtual and non-virtual settings (Gruss et al., 2015; Irazoqui et al., 2015; Zhang et al., 2014; Yarom and Falkner, 2014). Therefore, it can be performed in both environments, i.e. virtualisation and non-virtualisation. For instance, in cloud and virtual environments, by conducting the Flush+Reload Attack, the adversary will be able to exfiltrate GnuPG (a popular cryptography package that is utilised as the cryptography module of many open-source projects) private keys across several processor cores and virtual machines (Yarom and Falkner, 2014). Due to its generic nature, Flush+Reload Attack can be performed for other malicious purposes too. For instance, an attacker can launch a Flush+Reload Attack to gather statistical data on network traffic by observing network handling code or monitoring keyboard drivers to derive keystroke timing information.

Flush+Reload Attack consists of three stages (Zhang et al., 2014; Yarom and Falkner, 2014), consisting of Flush, Flush+Reload Interval and Reload. Stage one, Flush, involves flushing the observed memory line from the cache hierarchy including the shared last-level cache utilising clflush instruction (Gruss et al., 2015; Zhang et al., 2014; Yarom and Falkner, 2014). In stage two, Flush+Reload Interval, the attacker waits for a “prespecified interval” to enable the victim to access the memory line, while the last-level cache is employed by the victim running on the CPU core. Stage three, Reload, the attacker involves the attacker reloading the memory line, calculating the time to load it. A faster reload will indicate the existence of certain chunks in the last-level cache and the fact that they were run by the victim during the Flush+Reload interval. In contrast, a slower reload signifies the contrary (Zhang et al., 2014; Yarom and Falkner, 2014).

3.7. Asynchronous Attack

An Asynchronous Attack (AA) (Tromer et al., 2010) is a very complex attack that is difficult to detect. As such, it is likely to be carried out only by very advanced attackers. Despite the fact that an AA is similar to a Cache-Based Attack (CBA) in that both are performed against RSA for processors with simultaneous multithreading (Percival, 2005; Osvik et al., 2006), their cryptanalysis is very different because “algorithms and time scales involved in RSA vs. AES” executions are very different (Tromer et al., 2010). Under an AA, the adversary runs its own program on the same process as the encryption application (Tromer et al., 2010, Percival, 2005). This is achieved without any “explicit interaction” including “inter-process communication” (Crane et al., 2015; Gullasch et al., 2011; Osvik et al., 2006). The only insight that the attacker is required to have concerns the plaintexts or ciphertexts (as opposed to their exact values) (Acııçmez et al., 2007; Osvik et al., 2006; Percival, 2005). If the cipher runs on a simultaneous multi-threading (SMT) machine (Crane et al., 2015, Tromer et al., 2010), and the attacker is able to execute a dummy process concurrently with the cipher process (Gullasch et al., 2011), he will then be able to clear the BTB through the executions of the dummy process and causes a BTB miss during the operation of the target branch (Ge et al., 2016; Evtyushkin et al., 2016; Hund et al., 2013; Acııçmez et al., 2007' Hu, 1992). In this case, “the BPU automatically predicts the branch not to be taken if it misses the target address in the BTB”. Thus, there will be a
misprediction whenever the result of the target branch is “taken”. The attacker will be able to simulate the exponentiations partition of the sample according to the outcome of the branch (Acııçmez et al., 2007; Grunwald and Ghasi, 2002).

3.8. Evict+Time Attack

Evict+Time Attacks can be defined as a generic Cache-Timing Attack technique, by means of which the adversary activates multiple victim computations and calculates the victim’s runtime (Gruss, 2017; Irazoqui et al., 2014; Osvik et al., 2006). In order to do so, he evicts the cache set the victim’s runtime in order to calculate the effect of a particular cache set, and then and force the encryption programme to fetch key values from the main program (Crane et al., 2015; Tromer et al., 2010). In cases in which there is a timing difference when ejecting the cache set, the adversary will be able to deduce that the cache set was utilised by the victim computing process (Yarom and Falkner, 2014; Weiß et al., 2012; Acııçmez et al., 2007, Bernstein, 2005). To carry out an Evict+Time attack, the adversary will need to be able to calculate the precise starting and end time of a victim computing process. Evict+Time Attacks have been extensively covered in the literature. For instance, Osvik et al. (2006) examined Evict+Time Attacks in an attack on OpenSSL AES. Lipp et al. (2016) and Spreitzer and Plos (2013) illustrated that Evict+Time Attacks on OpenSSL AES are also relevant to mobile ARM-based devices. Similarly, Hund et al. (2013) showed that Evict+Time Attacks can be applied to compromise Kernel Address Space-Layout Randomisation (KASLR). Therefore, we do not aim to delve into this attack any further.

3.9. Timing Attacks Against Floating-Point Instructions

An attacker can also launch a TBSCA against the floating-point instructions of modern x86 processors (Andrysco et al., 2015; Coppens et al., 2009). The “running time of floating point addition and multiplication instructions can fluctuate by two orders of magnitude” (Andrysco et al., 2015; Hachez and Quisquater, 2000; Walter, 1999) depending on their operands (Coppens et al., 2009). Multiplying or dividing with subnormal values results in slowdown on, for instance, all tested Intel and AMD processors (Ge et al., 2016; Andrysco et al., 2015), irrespective of employing single instruction multiple data (SIMD) or x87 instructions (Intel®, 2016). Furthermore, by exploiting the aforementioned effects, an attacker will be able to employ the subnormal floating-point numbers to launch a timing attack, for instance, on a scalable vector graphics (SVG) filter, that reads arbitrary pixels from any victim web page though the Firefox browser as demonstrated by (Andrysco et al., 2015). Similarly, in relation to this type of exploit, two other researchers (Stone, 2013; Kotcher et al., 2013) have implemented a new method for cross-origin pixel stealing in the browser, which is a timing side-channel within CSS Scalable Vector Graphics (SVG) transforms. Such transforms can be employed through CSS to any element of a webpage, for instance iframes. Once the content of cross-origin is contained in an iframe, the containing page is then able to employ SVG transformation filters to that iframe (Andrysco et al., 2015). As a result, an attacker will be able to extract any pixel value from a website that he does not own by selecting certain SVG filters and determining the page render times.

3.10. Bernstein’s Attack

Bernstein’s Attack (Bernstein, 2005) is another variant of TBSCA that is carried out remotely on an AES T-table implementation in which the attacker can recover the AES key from “known-plaintext timings of a network server” on a different computer. This attack is the resultant of the fault in AES design and not to a specific library used by the server (Bernstein, 2005). Through this attack, Bernstein (2005) demonstrated that attacks as such were not restricted just to the Pentium III but instead could be performed against an “AMD Athlon, an Intel Pentium III, an Intel Pentium M, an IBM PowerPC RS64 IV, and a Sun UltraSPARC II”. By performing the Bernstein’s Attack, the adversary can potentially compromise T-table lookups in a system, that represent “pre-processed S-box computations” based on AES design (Gruss, 2017; Daemen and Rijmen, 2013). Through this attack,
the whole AES algorithm can be utilised as a fast sequence of T-table lookups that will be accessed based on an established implementation (Gruss, 2017; Bernstein, 2005). Such accesses can then be cached, and the timing difference can be monitored (Bernstein, 2005). The attacker will subsequently be able to determine which T-table entry was accessed by monitoring the timing difference. Many researchers (Spreitzer and Gérard, 2014; Weiß et al., 2014; Spreitzer and Plos, 2013; Neve et al., 2006; Bonneau and Mironov, 2006) have reproduced and assessed Bernstein’s Attack, that is presented in Bernstein’s (2005) seminal study.

3.11. Branch Prediction Attack

A Branch Prediction Attack (BPA) combined with cache performance can be a potential source of control-dependent and data-dependent timing (Coppens et al., 2009; Chen et al., 2003). As already stated, in modern processors, many resources are shared between different threads being run in the system (Ge et al., 2016; Coppens et al., 2009; Aciicmez et al., 2007). As a result, there will be conflict between those resources (Braun et al., 2015; Fedorova et al., 2010; Bonneau and Mironov, 2006). This leads to “inter-thread timing dependencies” (Ge et al., 2016), where the operation of one thread affects the timing performance of other threads (Martin et al., 2012). Therefore, an attacker will be able to monitor other threads competing for other resources to deduce information on condition that he does not have direct access to the timing of a thread which has come under attack (Coppens et al., 2009; Osvik et al., 2006; Bonneau and Mironov, 2006). To be able to launch a direct timing attack, the adversary needs to know the Branch Prediction Unit (BPU) architecture and also the implementation details of the encryption, as these two elements establish the prediction of the target branch (Aciicmez et al., 2007). Although this information is not easily available to the attacker, he can still carry out the examination stage speculating each possible state one at a time. The DTA can be used on any system on condition that branch prediction algorithm is applied on it. To compromise a cipher carrying out a BPA, the adversary requires having a result which must rely on the secret/private key of the cipher (Aciicmez et al., 2007). Furthermore, BPAs that are based on hardware performance can also be performed to elicit RSA keys from “exponentiations” executed in other processes (Gruss, 2017, Bhattacharya and Mukhopadhyay, 2015; Rebeiro et al., 2015).

3.12. Brief Overview of Timing-Based Attacks in Other Platforms

Although the focus of this study has been only on TBSCAs against PC platforms, nevertheless, we consider it worthwhile to provide a generic description of the way in which such attacks can be also carried out against other platforms such as mobile device. TBSCAs against mobile devices take advantage of both physical and software properties. For instance, a malign application can leverage the “accelerometer sensor” so as to launch an attack against the victim input. This is feasible because of the integral input technique that depends on touchscreens (Spreitzer et al., 2016; Aviv et al., 2012; Cai and Chen, 2011). Therefore, to perform a successful TBSCA against a mobile device, the adversary needs either to have a physical access to the device or remotely spread an application that appears to be benign (such as a game app) through an existing App store (Spreitzer et al., 2017; Spreitzer et al., 2016:). For example, through their study, O’Flynn (2016) illustrated that by shorting the “power supply of an off-the-shelf Android smartphone”, the attacker would be able to present a fault that can result in an invalid a fault loop count (Spreitzer et al., 2017). Attackers can also exploit the logical property of software provided by the API of the mobile device OS or even the OS itself (Spreitzer et al., 2016; Michalevsky et al., 2015; Zhou et al., 2013) to be able to carry out TBSCAs against such devices. This suggests that smartphones expand the extent of TBSCAs (Acar et al., 2016; O’Flynn, 2016; Spreitzer et al., 2016).

On the contrary, TBSCAs mounted against cloud computing hardware does not require the adversary to be in possession of the physical hardware (This, however, does not apply in cases where the cloud service provider, himself, is the adversary) since the attacker can potentially run a malicious application remotely (Spreitzer et al., 2016). For instance, to do so, he will require to be able to exploit
the Microarchitectural performance or software characteristics in order to be able to deduce private information from co-located processes (Ge et al., 2016; Spreitzer et al., 2016; Gruss et al., 2015; Yarom and Falkner, 2014; Tromer et al., 2010). Furthermore, in certain circumstances, TBSCAs can also be mounted through websites without the adversary relying on the victim to install an application. Furthermore, adversaries can also launch Side-Channel Attacks against mobile devices by taking advantage of the logical property of software that is offered by the API of the mobile device OS or even the OS itself (Spreitzer et al., 2016; Michalevsky et al., 2015; Zhou et al., 2013).

There exist other ways of launching TBSCAs in various platforms, the analysis of which is beyond the scope of this paper due to the page constraints.

4. SIDE-CHANNEL ATTACKS AGAINST RSA

Although many studies have been conducted on the topic of RSA, there are very few works that explore its vulnerabilities to TBSCAs. As a result, RSA’s vulnerabilities to TBSCAs are not still fully known by the research community. Therefore, we have assigned this section exclusively to the topic of RSA’s vulnerabilities to TBSCAs in an attempt to provide a more in-depth analysis of the principles underlying its susceptibilities to TBSCAs.

4.1. Overview of RSA Algorithm

RSA algorithm named after its creators, Rivest-Shamir-Adleman, (Rivest, 1978) is a public key encryption algorithm that is widely utilised to secure sensitive data transmission, especially when transmitted over an insecure network. RSA can be embedded in SSL (Secure Sockets Layer) to provide security and privacy over the Internet. In cryptography field, an asymmetric key algorithm employs a pair of different cryptographic keys to perform encryption and decryption. Both keys are mathematically connected, denoting that a message encrypted by the algorithm using one key can be decrypted by the same algorithm such as RSA. The RSA algorithm includes four parts: key generation, key distribution, encryption and decryption. The underlying fundamental behind RSA is the notion that it is applied to discover three big positive integers ‘e’, ‘d’ and ‘n’ in such a way that with modular exponentiation for all integer m (with 0 ≤ m < n and that even knowing e and n or even m), it can be very difficult to identify d:

\[ (m^e)^d \equiv m \pmod{n} \]

Furthermore, the RSA consists of both a public key and the private key. The public key allows the sender to carry out the encryption whereas the private key remains secret by the receiver and allows him to conduct the decryption. In more technical details, the public key, which contains the “modulus” n and the public “exponent” e, can be known by all parties and is applied to encrypt sensitive data. The idea behind the public key is that any data encoded with the public key can only be decoded in a sensible amount of time by employing the private key. Integer e and integer n, that are created by two main numbers p and q, represent the public key, and integer d represents the private key. However, integer d can also be utilised while decoding the data. This denotes that integer d can be regarded as a constituent of the private key too. In contrast, integer m represents the message. In contrast with the public key, the private key comprises the “modulus” n and the private “exponent” d, that must be reserved secret. The p, q, and n must also remain secret since they can be employed by the attacker to deduce d. There are two types of asymmetric encryption algorithms, both of which are built upon the Diffie-Hellman key agreement algorithm. Similarly, there are two unique types of symmetric key ciphers including block ciphers (fixed size) and stream ciphers (continuous stream). Although asymmetric encryption is much stronger and more secure than symmetric encryption, as
of yet, there does not exist an asymmetric key algorithm confirmed to be secure enough against a sophisticated mathematical attack. This is due to certain weaknesses in the asymmetric key algorithms that make them vulnerable to Timing Attacks. For instance, a savvy attacker can carefully measure the precise amount of time that it takes hardware to encrypt plaintext in order to facilitate the search for possible decrypting keys. Therefore, the usage of asymmetric key algorithms does not always guarantee security.

The potency of the RSA lies with the fact that it will be hard to factor large numbers. Almost, the most recognised factoring techniques are still considered to be slow as factoring large numbers can take many months and in some cases years. Since so many researchers have been attempting to factor large numbers in an efficient manner without any success, we can assume that at this point in time RSA is secure against factoring attacks for a standard n of 1024 bit in length (Wong, 2005). Securing the RSA against factoring attacks even more is possible by doubling the key length to 2048 bits or even more. Notwithstanding its potency against factoring technique attacks, RSA can be vulnerable to Timing Attacks. An adversary with an advanced knowledge of algorithms and programming skills could potentially exfiltrate RSA secret keys in a stealthy manner without directly breaching the RSA. Such an attack has come to be known as Timing Attack, which we have already covered in detail, that enables an adversary to monitor the execution time of a cryptographic algorithm and as a result infer the secret keys and values in the execution.

4.2. Timing Attacks Against RSA Algorithm

There are various types of covert and overt SCAs that can be mounted against RSA (Inci et al., 2016; Yarom and Falkner, 2014; Wang and Lee, 2006; Percival 2005), including Timing Attacks (the focus of this study) (Brumley and Boneh, 2005; Kocher, 1996), Trace-Driven Instruction Cache (Cai-Sen et al., 2011; Aciçmez, 2007) Differential Power Analysis (Bauer et al., 2013; Kocher et al., 1999), Electro Magnetic Emanations (Gandolfi et al., 2001; Genkin et al., 2015) and acoustic channels (Genkin et al., 2014; Hutter and Schmidt, 2013). Through a TBSCA against RSA, an attacker will be able to extract RSA private keys by utilising the Instruction Cache and the Branch Prediction Unit as covert channels (Inci et al., 2016; Aciçmez, 2007; Aciçmez et al., 2007). For instance, EL-Gamal (Zhang et al., 2012) private keys can be extracted from co-located VMs by taking advantage of emission in upper level caches. To launch a SCA against EL-Gamal, the attacker will require to use the Prime+Probe Attack in the LLC (Zhang et al., 2012; Liu et al., 2015). Likewise, the adversary can also employ the Branch Prediction Performance Counters to extract RSA keys (Inci et al., 2016; Bhattacharya and Mukhopadhyay, 2015). Furthermore, advanced attackers will also be able to extract RSA private keys from co-located VMs (Yarom and Falkner, 2014; Wang and Lee, 2006) by mounting the Flush+Reload Attack while memory De-Duplication is active (Yarom and Falkner, 2014).

A TBSCA against RSA can be carried out if the adversary knows the victim’s hardware in adequate detail and if he is mathematically advanced enough to calculate the decryption time for multiple known ciphertexts. In such a case, the attacker will then be able to infer the decryption key, i.e. d, immediately (Kocher, 1999). Such an attack can also be performed against the RSA scheme (Brumley and Boneh, 2005) to extract RSA factorisation over a network connection from a web server enabled with Secure Sockets Layer (SSL). Another way of launching a TBSCA against RSA is through the exploitation of information that has been leaked by the Chinese Remainder Theorems Optimisation (a technique that is often utilised by many RSA implementations).

As already discussed, the Evict+Time Attack (Aciçmez et al., 2007) against RSA is performed to evict entries from the BTB in a selective manner by running branches at relevant addresses, and then monitor the impact on the runtime of an RSA encryption in OpenSSL. To do so, the attacker will need to calculate the time that is needed to carry out the initial ejection to be able to deduce whether OpenSSL had beforehand run the branch or not. Although the Simple Branch Prediction Analysis (SBPA) (Aciçmez et al., 2007) uses the same approach as that of the aforementioned approach, it is more powerful as it can exfiltrate most key bits in a single RSA execution. The SBPA is based
on the “fine-grained sharing” that is involved in SMT (Ge et al., 2016). Furthermore, attackers can identify and take advantage of the contention between hardware threads on the multiplier unit (Ge et al., 2016; Aciciosmez and Seifert’s, 2007; Wang and Lee, 2006). Simultaneous Multi-Threaed (SMT) processors (Tulsen et al., 1995), a method for optimising the overall efficiency of CPUs with hardware multithreading, execute multiple independent processes in parallel in order to make a more effective use of the resources offered by processor designs. The parallel threads share a cache of functional units (FUs) dynamically assigned to each process (Wang and Lee, 2006). Such a sharing creates interference between two difference processes as they will need to compete for FUs. This results in a ‘covert channel’, the extent of which surpasses those of other covert channels. Similarly, the conflicts between hardware threads on the multiplier unit can create a ‘side channel’ that enables a malign thread to differentiate multiplications from “squaring in OpenSSL’s RSA architecture” (Ge et al., 2016; Aciciosmez and Seifert’s, 2007; Wang and Lee, 2006). These aforementioned attacks can enable an adversary to calculate the latency created by contending threats that are compelled to remain until they can access the multiplier unit.

The scatter-gather implementation employed in the modular exponentiation routine in OpenSSL is also prone to the CacheBleed attack, in which cache-bank conflicts on the Sandy Bridge microarchitecture can be exploited by the attackers (Fog, 2017; Intel®, 2016; Yarom et al., 2017). CacheBleed Attack, which has successfully been tested on an Intel Xeon E5-2430 processor (Yarom et al., 2017), allows the attacker to detect the cache pool that maintains each given multiplier utilised through the “exponentiation in the OpenSSL constant time RSA design” (Aciciozmé et al., 2007; Brickell et al., 2006). As a result, it enables the attacker to exfiltrate the entire private key after he has monitored 16,000 decryptions for 4096-bit RSA (Yarom et al., 2017, Ge et al., 2016).

Furthermore, a thread that is running on a design of a SMT processor is also vulnerable to denial of service through a malign thread. This results in a significant reduction in the speed of the original thread. Therefore, an adversary can utilise Performance Counter Hardware to create this type of slowdown by intentionally abusing the shared resources and design decisions that are essential for high speed implementation (Grunwald and Ghiasi, 2002). Consequently, since a given thread can deny other threads (in resource sharing) of their resources through the usage of a multithreading processor, one thread can have an impact on the performance of another thread. Applying exceptional conditions on behalf of one thread can also create a significant performance degradation for another SMT thread (Ge et al., 2016; Grunwald and Ghiasi, 2002). Moreover, in certain processors (such as the Intel Pentium 4), self-adaptive code flushes the trace cache causing a significant reduction in performance (e.g. in a DoS attack). Although control techniques facilitated by resource sharing are capable of enhancing essential processor speed-paths, they can be taken advantage of by a single action by a malicious thread that can create many sets of delays.

Another type of Timing Attack against RSA is “Square vs. Multiplication” that is relevant to parallel multi-threading CPU designs. This attack cannot be carried out on CPU implementations without the aid of SMT hardware. Unlike other aforementioned attacks against RSA that often focus on the notion of a shared resource, Square vs. Multiplication attack is based on the concept that Intel’s hyper-threading technology shares the ALU’s large floating-point multiplier among its two hardware threads (Aciciosmez and Seifert, 2007). An attacker can also exploit the timing performance of the Montgomery multiplications during the initialisation of the table enabling himself to extract one of the main factors of timing RSA moduli by adding to the number of multiplications (Biswas et al., 2017; Aciciozmé et al., 2005; Brumley and Tuveri, 2011; Brumley and Boneh, 2005). In addition, by performing a TBSCA, the entire RSA key in cloud settings can be recovered (Inci et al., 2016; Irazoqui et al., 2014; Yarom and Falkner, 2014) by the attackers, outside the closed-box VM, who can potentially exploit a vulnerable Virtual Machine Monitor (VMM) system that is running on top of a SMT processor (Inci et al., 2016; Irazoqui et al., 2014). A TBSCA on a hyper-threading processor also allows an attacker through a user processor to deduce the RSA key of another processor which is carrying out RSA encryption (Wang and Lee, 2006; Percival, 2005). There will be no need for
any type of special tools to facilitate this attack, and the attack also does not even need to rely on software defects to extract RSA. Only a series of memory accesses are required to be executed by the spy process, followed by its observation of the timing while the victim process is being executed on the same processor (Osvik et al., 2006; Wang and Lee, 2006; Bernstein, 2005; Percival, 2005).

5. CONCLUSION

In this study, we identified and analysed some of the existing known Timing-Based Side-Channel Attacks (TBSCAs), and demonstrated their devastating impacts on shared, modern computing hardware. We thoroughly reviewed relevant literature within the context and we discussed various attack vectors that attackers can adopt to mount such attacks against components of modern PC platforms. Through this systematic literature review and analysis, one can deduce that all Microarchitectural Timing Attacks, irrespective of their type, can exploit security systems, regardless of advanced partitioning methods (e.g. memory protection), sandboxing or even virtualisation. Hence, it is vital to identify every conceivable Microarchitectural susceptibility in order to comprehend the potential of Microarchitectural analysis and design to implement more secure systems. Although this study mainly focused on the review and analysis of Timing Attack vectors, in a follow-up paper as a future work, we are providing the existing countermeasures against such attacks and propose new strategies to deal with these attacks. There are already comprehensive research works (Yang et al., 2018; Sohal et al., 2018; Kuo et al., 2018) covering the response to such attacks; we will discuss these as a future work of this piece of research.
REFERENCES


Irazoqui, G., Inci, M. S., Eisenbarth, T., & Sunar, B. (2014). Wait a Minute! A Fast, Cross-VM Attack on AES. In International Workshop on Recent Advances in Intrusion Detection (pp. 299-319). doi:10.1007/978-3-319-11379-1_15


---

Reza Montasari is a lecturer at the School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University. He holds a BSc (Hons) in Multimedia Computing, an MSc and PhD both in Digital Forensics. He is also a member of IET (MIET) and is currently working towards becoming a CEng. Reza has published numerous research papers and serves as a programme committee member for various reputable international journals and conferences.

Amin Hosseinian-Far holds the position of Senior Lecturer in Business Systems & Operations at the University of Northampton, United Kingdom. In his previous teaching experience, Amin was a Staff Tutor at the Open University, UK, a Senior Lecturer and Course Leader at Leeds Beckett University. He has held lecturing and research positions at the University of East London, and at a number of private HE institutions and strategy research firms. Dr. Hosseinian-Far has also worked as Deputy Director of Studies at a large private higher education institute in London. He received his BSc (Hons) in Business Information Systems from the University of East London, an MSc degree in Satellite Communications and Space Systems from the University of Sussex, a Postgraduate Certificate in Research and a PhD degree titled ‘A Systemic Approach to an Enhanced Model for Sustainability’ which he acquired from the University of East London. Dr. Hosseinian-Far holds Membership of the Institution of Engineering and Technology (IET), Senior Fellowship of the Higher Education Academy (HEA), and Fellowship of the Royal Society of Arts (RSA).

Richard Hill is Head of the Department of Computer Science and Director of the Centre for Industrial Analytics at the University of Huddersfield, UK. Professor Hill has published widely in the areas of Big Data, predictive analytics, the Internet of Things, edge analytics and Industry 4.0, and has specific interests in digital manufacturing.