# Cache-Timing Attacks on RSA Key Generation

Alejandro Cabrera Aldaya[1], Cesar Pereida García[2],
Luis Manuel Alvarez Tapia[1] and Billy Bob Brumley[2]

[1] Universidad Tecnológica de la Habana (CUJAE), Habana, Cuba
{aldaya,lalvarezt89}@gmail.com
[2] Tampere University, Tampere, Finland
{cesar.pereidagarcia,billy.brumley}@tuni.fi

**Abstract.** During the last decade, constant-time cryptographic software has quickly transitioned from an academic construct to a concrete security requirement for real-world libraries. Most of OpenSSL's constant-time code paths are driven by cryptosystem implementations enabling a dedicated flag at runtime. This process is perilous, with several examples emerging in the past few years of the flag either not being set or software defects directly mishandling the flag. In this work, we propose a methodology to analyze security-critical software for side-channel insecure code path traversal. Applying our methodology to OpenSSL, we identify three new code paths during RSA key generation that potentially leak critical algorithm state. Exploiting one of these leaks, we design, implement, and mount a single trace cache-timing attack on the GCD computation step. We overcome several hurdles in the process, including but not limited to: (1) granularity issues due to word-size operands to the GCD function; (2) bulk processing of desynchronized trace data; (3) non-trivial error rate during information extraction; and (4) limited high-confidence information on the modulus factors. Formulating lattice problem instances after obtaining and processing this limited information, our attack achieves roughly a 27% success rate for key recovery using the empirical data from 10K trials.

**Keywords:** applied cryptography · public key cryptography · RSA · side-channel analysis · timing attacks · cache-timing attacks · OpenSSL · CVE-2018-0737

## 1 Introduction

Side-channel analysis (SCA) continues to be a serious threat against the security of systems and cryptography libraries. Specifically, microarchitecture attacks and cache-timing attacks are gaining more traction due to the severe architecture flaws recently discovered in many microprocessors [Koc+19, Lip+18]. Cache-timing attacks are attractive for attackers and researchers due to the ability to perform them semi-remotely and without special privileges. So far, practical cache-timing attacks have been developed against multiple cryptosystems, including but not limited to DSA [PGBY16], ECDSA [PGB17], DH [GVY17] and RSA [YGH16]. As a countermeasure against this type of attack, cryptography library developers such as OpenSSL and forks integrate algorithms in their codebase that execute in constant-time independently of the input values. During recent years, several researchers discovered and exploited flaws in these mitigations.

SCA research focuses mainly on cryptographic operations such as encryption, decryption, key exchange, and signature generation. All of them have in common the repeated use of the private key as input during some step of the algorithm execution, thus being able to observe and capture the leakage over several runs. In contrast, SCA research targeting key generation seems to be neglected—we speculate due to two assumptions: (1) keys are only generated once during the initial stage in a secure environment, isolated from any

possible threats; and (2) single trace attacks pose too many challenges, e.g. noise, and are not feasible.

We motivate our work with a scenario where a malicious attacker is co-located with a victim process generating RSA keys. In services such as Let's Encrypt[1], RSA key generation is a common, regular and semi-predictable operation for web server automated certificate renewal, often performed in shared cloud environments such as Amazon Web Services (AWS), and Microsoft Azure. Recent numbers reported by Censys[2] suggest Let's Encrypt is now the largest certificate issuer, therefore generating thousands of keys, and with an adoption rate of more than 60% from websites using SSL/TLS, thus highlighting the need for SCA-hardened key generation.

In this work, we present a methodology developed to identify the use of known side-channel vulnerable functions in cryptography libraries such as OpenSSL. Using our methodology, we disclose several vulnerabilities affecting the OpenSSL RSA key generation implementation. Due to the impact of our attack, the OpenSSL security team issued a security advisory and CVE-2018-0737. Moreover, we present the first practical single trace cache-timing attack against the binary GCD step used during RSA key generation leading to complete RSA private key recovery. The root cause of the vulnerability is the GCD callee function not supporting the constant-time flag, compounded by the parent function's failure to enable it. More precisely, our attack focuses on the execution of the non constant-time binary GCD algorithm to test the coprimality between the integers $p-1$ and $q-1$, and the public exponent $e$. Finally, this work serves as a reminder that cryptography libraries should strive for a secure by default approach, thus avoiding several side-channel attacks that still might be lurking in the codebase.

Our contributions in this work are the following: (1) We develop a methodology to identify insecure code paths through known side-channel vulnerable functions still in use by cryptography libraries, and use it to identify and exploit a flaw in OpenSSL that allows a practical single trace cache-timing attack against RSA key generation (Section 3.1); (2) We combine several techniques from cache-timing attacks and power analysis to capture traces during binary GCD execution and process them in order to obtain a sequence of shift and subtraction operations, i.e. algorithm state, related with prime values $p$ and $q$ (Section 3.4); (3) Building on existing RSA key recovery work, we propose a novel error correction algorithm for noisy RSA primes that allows us to recover roughly 50% of bits for each prime (Section 4); (4) We implement a lattice attack that factors RSA-2048 keys knowing 522 bits of one prime. We perform an end-to-end attack for 10K independent keys achieving roughly a 27% success rate, with room for improvement (Section 5).

# 2 Background

## 2.1 The RSA Cryptosystem

RSA is a public key cryptosystem invented in 1978 [RSA78]. An RSA public key is a tuple of integers $(N, e)$ where $p$ and $q$ are primes and $N = pq$ holds, and furthermore $ed = 1 \bmod (p-1)(q-1)$ holds, implying both $e$ and $d$ are odd. For the remainder of this paper, we restrict to standardized RSA-$n$ that mandates for $n$-bit $N$ both $p$ and $q$ have bit-length $n/2$ and furthermore $2^{16} < e < 2^{256}$ holds [Fip]. For efficiency reasons, $e = 65537$ is the most common choice.

The private key is the tuple $\mathtt{sk} = (p, q, d, dp, dq, iq)$ where the latter three are Chinese Remainder Theorem (CRT) values not relevant to this work. For well-chosen parameters, recovering the private key from the public key is believed to be as hard as factoring

[1] https://letsencrypt.org/

[2] https://censys.io/certificates/report?q=tags%3Atrusted&field=parsed.issuer.organization.raw&max_buckets=50

$N$ [Hin10]. Regarding security, the current minimum recommended RSA key size is 2048 bits, implying that $p$ and $q$ are 1024-bit primes. Applications of RSA in cryptography include public key encryption and digital signatures.

The secrecy of $p$ and $q$ is essential for RSA, moreover, partial knowledge of either value can lead to polynomial-time factoring algorithms. In his groundbreaking work, Coppersmith [Cop96] proved that knowing half of the bits from one prime suffices to factor $N$ in polynomial time, a critical point in our attack (see Section 5).

**Algorithm 1:** OpenSSL RSA key generation

**Input:** Key size $n$ and public exponent $e$.
**Output:** Public and private key pair.

1 **begin**
2  **while** $\gcd(p-1, e) \neq 1$ **do**
3  $p \leftarrow$ rand $n/2$-bit prime  `/* Generate p */`
4  **while** $\gcd(q-1, e) \neq 1$ **do**
5  $q \leftarrow$ rand $n/2$-bit prime  `/* Generate q */`
6  $d \leftarrow e^{-1} \bmod (p-1)(q-1)$  `/* Priv exp */`
7  $dp \leftarrow d \bmod (p-1)$
8  $dq \leftarrow d \bmod (q-1)$  `/* CRT parameters */`
9  $iq \leftarrow q^{-1} \bmod p$
10  **return** *(N, e), (d, p, q, dp, dq, iq)*

OpenSSL's RSA key generation closely resembles Algorithm 1. The first steps aim at generating random secret primes $p$ and $q$, during which two loops ensure that $(p-1)$ and $(q-1)$ are coprime with $e$. Steps 7-9 are not relevant to this work.

Algorithm 1 involves computing (at least) two GCDs and two modular inversions. Binary GCD algorithms are a common implementation choice for both of these operations—a description follows.

## 2.2 Binary GCD Algorithms

Stein [Ste67] proposed the binary greatest common divisor algorithm (binary GCD) in 1967. This algorithm computes the GCD of two integers $a$ and $b$ employing only right-shift operations and subtractions (Algorithm 2). This approach is very attractive in cryptography as it performs very well, especially with large inputs.

In OpenSSL, GCD computations use the function `BN_gcd`, a high level wrapper to the function `euclid` that is one implementation of the binary GCD. Note however, that it does not follow the *classic* algorithm structure (c.f. Algorithm 2). Regardless, its flow can be analyzed using the classic variant since their equivalence can be easily verified. If an adversary can distinguish a right-shift from a subtraction operation, the algorithm state can be recovered [AGS07, ACSS17, PGB17]. We expand on this concept later in this section.

Finally, it is worth noting that with respect to RSA key generation, Step 4 never executes since one of the inputs is always odd.

## 2.3 Binary GCD: Side-Channel Analysis

The execution flow of Algorithm 2 is highly dependent of its inputs. In Algorithm 2 some execution flow relevant steps are highlighted. The *u-loop* and *v-loop* are the loops that remove all power-of-two divisors in variables $u$ and $v$ at each iteration. The *sub-step* executes when both variables are odd, consisting of a single subtraction.

**Algorithm 2:** Binary GCD

**Input:** Integers $a$ and $b$ such that $0 < a < b$.
**Output:** Greatest common divisor of $a$ and $b$.

```
1  begin
2      u ← a, v ← b, i ← 0
3      while even(u) and even(v) do
4          u ← u/2, v ← v/2, i ← i + 1
5      while u ≠ 0 do
6          while even(u) do
7              u ← u/2                                  /* u-loop */
8          while even(v) do
9              v ← v/2                                  /* v-loop */
10         if u ≥ v then
11             u ← u − v                                /* sub-step */
12         else
13             v ← v − u
14     return v · 2^i
```

Consistent with the existing literature [ACSS17, PGB17], we encode the execution flow sequence of this algorithm with two symbols 'L' and 'S' representing right-shift and subtraction, respectively. Another representation uses two variables $Z_i$ and $X_i$ defined[3] in [ACSS17] as follows: (1) $Z_i$ stores the number of right-shifts at iteration $i$. (2) $X_i$ stores a binary value to represent the result of the condition (Step 10 in Algorithm 2) at iteration $i$. $X_i$='$u$' means the condition was true while $X_i$='$v$' the opposite.

Figure 1 shows an LS-sequence example of an execution flow. The sequence reads from left to right: $Z_0 = 3$, $Z_1 = 2$, $Z_2 = 1$, $Z_3 = 5$, etc.

LLLSLLSLSLLLLLSL...LS

**Figure 1:** LS-encoded binary GCD execution flow example.

Regarding SCA, there are three different models for analyzing Algorithm 2 leakage. Each model originally targets the Binary Extended Euclidean Algorithm (BEEA) for computing modular inverses. However, they also apply to Algorithm 2 because the models exploit the execution flow leakage w.r.t. variables $u$ and $v$, and said flow is the same for both algorithms when executed with the same input pair.

**All-or-nothing.** Acıiçmez, Gueron, and Seifert [AGS07] and Aravamuthan and Thumparthy [AT07] independently proposed this model in 2007. It requires that the adversary knows *all* $Z_i$ and $X_i$ to recover algorithm inputs.

**Partial.** Aldaya, Cabrera Sarmiento, and Sánchez-Solano [ACSS17] recently proposed this model, an algebraic approach that relates the number of known $Z_i$, $X_i$ with the number of input bits that can be recovered. In comparison with the previous model, this approach is more flexible as it can extract information from partial knowledge of the execution flow. In this model, the number of recovered bits grows with the number of $Z_i$ and $X_i$

[3] Equivalent to SHIFTS[i] and SUBS[i] definition by Acıiçmez, Gueron, and Seifert [AGS07]

that an adversary knows. Our work employs this model (see Section 3.2 for this selection rationale).

**Look-up.** Pereida García and Brumley [PGB17] proposed a model that also allows partial recovery, but instead of an algebraic approach it employs a table look-up. The adversary generates a table that relates every LS-sequence of a given length with the corresponding partial input bits. This model performs better than the previous when the number of bits to recover is small as it captures some algebraic equivalences not previously modeled. However, it becomes impractical for recovering a large number of bits (i.e. longer LS-sequences). The computational complexity (time and storage) for creating a table containing all possible LS-sequences increases exponentially on the LS-sequence length.

## 2.4 The Flush+Reload Technique

This technique is a cache-based side-channel attack technique targeting the Last-Level Cache (LLC) and used during our attack. Flush+Reload is a high resolution, high accuracy and high signal-to-noise ratio technique that positively identifies accesses to specific memory lines. It relies on cache sharing between processes, typically achieved through the use of shared libraries or page de-duplication.

A round of attack consists of three phases: (1) The attacker evicts (i.e. flushes) the target memory line from the cache. (2) The attacker waits some time so the victim has an opportunity to access the memory line. (3) The attacker measures the time it takes to reload the memory line. The latency measured in the last step tells whether or not the memory line was accessed by the victim during the second step of the attack, i.e. identifies cache-hits and cache-misses, in addition to information about the cache activity for detecting spy preemptions.

Consult [YF14, All+16, PGBY16] for more information on the technique and discussions about the challenges during attack setup due to processor optimizations and different architectures.

## 2.5 Error Correction in RSA Keys

Following Section 2.1, an RSA private key is composed by a six element tuple $\mathtt{sk} = (p, q, d, dp, dq, iq)$. The knowledge of any of these elements directly allows breaking the system. In addition, these elements contain redundancy and are related through public information. For example, as $N = pq$, the relation $N \equiv pq \bmod 2^n$ holds for every positive integer $n$. Therefore the knowledge of the first $n$ bits of $p$ directly reveals the same amount on $q$.

Motivated by some implementation attacks such as side-channel and cold-boot attacks, factoring $N$ from a noisy version of $\mathtt{sk}$ is an active research line since the pioneering works of Percival [Per05] and Heninger and Shacham [HS09].

The number of elements in a noisy version of $\mathtt{sk}$ depends on the data leak. For example, Percival [Per05] assumes a noisy version of $(dp, dq)$ whereas Heninger and Shacham [HS09] consider different versions of noisy $\mathtt{sk}$, such as $(p, q, d, dp, dq)$ and $(p, q)$. The number of elements in the noisy $\mathtt{sk}$ is often denoted in literature as $m$.

Regarding this work, the case of $m = 2$ is particularly interesting since a Flush+Reload attack allows an attacker to obtain a pair of noisy LS sequences related to $p$ and $q$, therefore we focus further related work analysis on the $m = 2$ case.

Another implementation attack property is the nature of errors that produces the noisy $\mathtt{sk}$—termed *noise model* by Henecka, May, and Meurer [HMM10]. They defined the noise model of Percival [Per05] and Heninger and Shacham [HS09] works as the *erasure model*. In these works, the adversary knows which bit positions contain valid data, and likewise exactly at which bit position data is missing.

On the other hand, Henecka, May, and Meurer [HMM10] proposed an algorithm for correcting errors in RSA keys from noisy `sk` where some bits are flipped. Correcting errors in this *bit-flip model* is more challenging than in the erasure model [HMM10], however the authors showed that for $m = 2$ it is possible to achieve a success rate of 24% if the probability of a single bit-flip is less than 0.084. Paterson, Polychroniadou, and Sibborn [PPS12] also studied this error model but considering that a flip $0 \to 1$ has a different probability than $1 \to 0$.

Kunihiro [Kun15] analyzed *erasure* and *bit-flip* models in a combined approach proposing an algorithm and a theoretical analysis of the bounds on erasure and bit-flip rates to succeed, improving the allowed bit-flip rate up to 0.11 in a scenario without erasures. For a detailed survey on these approaches we suggest the reader to consult [Kun18].

In addition to these error correction algorithms, some authors employ multiple traces to remove noise from `sk`. For example, in very different attack scenarios and completely independent research, Irazoqui, Eisenbarth, and Sunar [IES16] and Schwarz et al. [Sch+17] obtained an error rate of no more than 0.04 from a single trace attack. Then, combining the same data leakage from five traces, they corrected all the errors in their respective setting. RSA key error correction using multiple traces is an approach that only applies when the attacker can capture multiple traces of the same operation leaking data.

## 2.6 Related Work

**Attacks on RSA keys.** Over the years, cryptanalysis of RSA keys has been performed due to its widespread usage, its mathematical structure (i.e. CRT-based methods) and the ease of generating low entropy keys. One classification of attacks against RSA keys is: (1) only public key knowledge; (2) partial private key knowledge [Hin10].

The first category assumes an attacker only has knowledge of the public key $(N, e)$, attempting to use factoring methods such as Pollard $p - 1$ [Pol74], Pollard Rho [Pol75] and sieving methods to recover the private factors $p$ and $q$. This type of attack is bound by the often sub-exponential, yet intractable, time complexity of the factoring methods, requiring massive computation time and resources. Current research achieves factorization of 768-bit RSA keys [Kle+10], therefore it has limited practical applicability and interest for an attacker.

The second category exploits partial knowledge about the private *and* public keys to perform attacks such as low exponent attacks [BM03, Wie90], side-channel attacks [YGH16, Bau+14], and Coppersmith related attacks [Cop96, Cop97], considered a universal tool to attack RSA keys with poorly chosen parameters or keys generated with poor entropy, i.e. using a faulty implementation.

In 2012, two independent teams [Hen+12, Len+12] exploited poor entropy of RSA keys in SSL certificates, SSH host keys, and PGP keys, thus allowing them to trivially factorize keys by carrying out pairwise GCD computations to recover shared prime factors among other RSA keys. Similarly in 2013, Bernstein et al. [Ber+13] analyzed the public record of RSA keys in the "Citizen Digital Certificate" database of Taiwanese citizens. The authors recovered 265 private keys by running a batch GCD computation followed by Coppersmith's method.

In 2017, Nemec et al. [Nem+17] discovered a critical vulnerability in the library used to generate RSA keys for identity cards, passports and Trusted Platform Modules; allowing factorization of 1024 and 2048-bit keys. Once again, this exploit was possible due to poor entropy introduced by a special mathematical structure of the prime factors that not only allowed key recovery using Coppersmith's method but also detection of keys with this special structure.

**Microarchitecture attacks on RSA.** In his seminal work, Percival [Per05] demonstrated a cache-timing attack against RSA by identifying access to precomputed multipliers stored

in memory when using the Sliding Window Exponentiation (SWE) algorithm implemented in OpenSSL version 0.9.7c. To mitigate this issue, the OpenSSL team added a "constant-time" implementation of the modular exponentiation algorithm combining a fixed-window exponentiation algorithm with a scatter-gather method [Bri+06], allowing to mask table access to the multipliers. The scatter-gather method ensures the same cache lines are always accessed, irrespective of the multiplier used.

In 2016, Yarom, Genkin, and Heninger [YGH16] showed that the previous scatter-gather method implemented in OpenSSL still leaked timing information. In their work, the authors exploited cache-bank conflicts by accessing the same offsets within a cache line, these offsets depend on the multipliers used which are decided based on the private key. The attack allows 4096-bit RSA key recovery after observing 16000 decryptions on a HyperThreading architecture.

More recently, Bernstein et al. [Ber+17] performed 1024 and 2048-bit key recovery in the Libgcrypt library when computing modular exponentiations using the left-to-right sliding window method. More precisely, the authors demonstrated that the direction of the sliding window matters since it leaks more or less information depending on the encoding direction. Applying the Flush+Reload technique, paired with the algorithm by Heninger and Shacham [HS09], the authors are able to efficiently reconstruct private keys using a side-channel leak after recovering roughly 50% of the secret bits.

Acıiçmez, Gueron, and Seifert [AGS07] showed information leakage in OpenSSL 0.9.8a during the modular inversion operation. When using the BEEA for modular inversion during key generation, decryption, and blinding when employing the RSA-CRT variant, these algorithms compute on secret values. Developing Simple Branch Prediction Analysis (SBPA), the authors conjecture it is feasible to deduce the outcome of branch statements using timings, recovering critical BEEA algorithm state, therefore leading to secret key recovery.

**Side-channel attacks on RSA key generation.** The research available on SCA against RSA key generation is limited and mostly focuses on leakages in physical devices. Finke, Gebhardt, and Schindler [FGS09] performed an attack on a custom implementation of a prime generation algorithm used for RSA key generation, analyzed using Simple Power Analysis (SPA). In 2012, Vuillaume, Endo, and Wooderson [VEW12] presented a Differential Power Analysis (DPA) template attack and fault attack on the Fermat and Miller-Rabin tests on a secure microcontroller but the authors give no additional information regarding their setup. Later on, Bauer et al. [Bau+14] analyzed the security of prime generation algorithms and the sieving process. Targeting the divisibility phase, the authors obtained more than half of the bits from the prime number generated with their own implementation and then using Coppersmith's technique they recovered 1024-bit RSA keys. More recently, Aldaya et al. [Ald+17] analyzed the modular inversion operation used during RSA private key generation, leading to full key recovery using SPA. This attack differs from previous works because it focuses on alternative routines invoked during key generation, instead of primality tests or prime number generation.

Moreover, recent independent work examines one of the three code paths analyzed in this work (i.e. the `BN_gcd` function). Weiser, Spreitzer, and Bodner [WSB18] target RSA key generation within an Intel SGX enclave by a noiseless controlled-channel page-fault attack. Controlled-channel attacks [XCP15] are privileged attacks originating from a malicious OS targeting SGX enclaves, aligned with the SGX threat model. The most important differences compared to our work are: (1) cache-timing attacks are unprivileged and do not require escalation to kernel space (i.e. a malicious OS); and (2) controlled-channels are error-free, while cache-timing channels are far from that.

# 3 RSA Key Generation: New Vulnerabilities

Originally introduced with OpenSSL 0.9.7 in 2005 following [Per05], the *constant-time flag* is a boolean for BIGNUM variables handling secret information such as private keys, secret prime values, nonces, and integer scalars. When the flag is set, and the executing algorithm supports the flag, the code takes an early exit to the constant-time version of the algorithm, otherwise continues executing the default insecure version. For the sake of performance, OpenSSL defaults to non constant-time functions, assuming most operations are not secret.

## 3.1 Insecure Code Paths: A Methodology

Two recent works exploit the insecure default behavior of OpenSSL's constant-time flag. Pereida García, Brumley, and Yarom [PGBY16] exploit the fact that, by design, the flag does not propagate from the source to the destination during BIGNUM copy operations. As a result, modular exponentiations during DSA sign operations took a side-channel insecure modular exponentiation path. Pereida García and Brumley [PGB17] exploit the failure to set the flag during ECDSA sign operations. In that case, the resulting scalar multiplication function is oblivious to the flag and always followed a side-channel secure path; the modular inversion function, however, requires this flag to follow its side-channel secure path.

These examples demonstrate that constant-time flag handling is tricky, hence there could be other vulnerable code paths believed to be safe. For tackling the problem of detecting such potential vulnerable code paths we developed a semi-automated tool that revises, with a single execution, multiple code paths, producing a report about them. Our methodology consists of the following steps: (1) From existing work, we create a list of previously known side-channel vulnerable functions within a library. (Here, OpenSSL.) (2) The tool utilizes the debugger to automatically set break points at lines of code, identified in the previous step, which should not be reached during security-critical operations. (3) The tool runs several security-critical commands and generates a report for calls hitting said break points.

Cryptography libraries such as OpenSSL support multiple architectures, compilation options, and implementations of the same functionality. Therefore, our tooling allows to perform exhaustive testing on the library, trying several combinations for vulnerable paths and easing the workload for SCA.

Using our tooling, w.r.t. RSA key generation we identified the following subset of known side-channel vulnerable functions of interest: (1) The function `BN_gcd` contains highly input-dependent branches that can potentially be used as a side-channel attack vector. Since the code has no early exit to a side-channel secure code path, i.e. does not check the constant-time flag at all, we blacklist the function's entry point. (2) The function `BN_mod_inverse` executes a check for the constant-time flag at the beginning of the function, and early exits to a side-channel secure path if it is set. If the flag is not set, it continues to a side-channel insecure path. We blacklist the line immediately following the early exit. (3) The function `BN_mod_exp_mont` is analogous to the above, yet for modular exponentiation. Similarly, we blacklist the line immediately following the early exit.

Figure 2 shows a visualization of our tooling, setting breakpoints on `bn_gcd.c` (Lines 120 and 238) and `bn_exp.c` (Line 418) based on the previously blacklisted lines. Our tooling executes OpenSSL `genpkey` command to generate an RSA key, hitting the three break points multiple times, and reporting their corresponding call stacks. Naturally, hitting the break points does not guarantee a vulnerability—a deeper analysis follows.

```
INFO: Parsing source code at: ./openssl-1.0.2k
...
INFO: Breakpoints file generated: triggers.gdb
...
INFO: Monitor target command line
TOOL: gdb --batch --command=triggers.gdb --args
      openssl-1.0.2k/apps/openssl genpkey -algorithm RSA
      -out private_key.pem -pkeyopt rsa_keygen_bits:2048
...
INFO: Setting breakpoints...
Breakpoint 1 at ...: file bn_exp.c, line 418.
Breakpoint 2 at ...: file bn_gcd.c, line 120.
Breakpoint 3 at ...: file bn_gcd.c, line 238.
...
INFO: Insecure code executed!
Breakpoint 1, BN_mod_exp_mont (...) at bn_exp.c:418
418    bn_check_top(a);
#0  BN_mod_exp_mont (...) at bn_exp.c:418
#1  ... in witness (...) at bn_prime.c:356
#2    ... in BN_is_prime_fasttest_ex (...) at bn_prime.c:329
#3    ... in BN_generate_prime_ex (...) at bn_prime.c:199
#4    ... in rsa_builtin_keygen (...) at rsa_gen.c:150
...
INFO: Insecure code executed!
Breakpoint 2, BN_gcd (...) at bn_gcd.c:120
120        int ret = 0;
#0    BN_gcd (...) at bn_gcd.c:120
#1    ... in rsa_builtin_keygen (...) at rsa_gen.c:154
...
INFO: Insecure code executed!
Breakpoint 3, BN_mod_inverse (...) at bn_gcd.c:238
238        bn_check_top(a);
#0    BN_mod_inverse (...) at bn_gcd.c:238
#1    ... in BN_MONT_CTX_set (...) at bn_mont.c:450
#2    ... in BN_is_prime_fasttest_ex (...) at bn_prime.c:319
#3    ... in BN_generate_prime_ex (...) at bn_prime.c:199
#4    ... in rsa_builtin_keygen (...) at rsa_gen.c:171
...
```

**Figure 2:** Testing the proposed methodology tool.

**Insecure exponentiation code path.** The Miller-Rabin primality test [Rab80] is the most common implementation of Algorithm 1, Lines 3 and 5. It involves choosing a random "witness" base $b$ then computing $b^x \bmod p$ where $p$ is the candidate prime and the relation $2^k x = p - 1$ holds. Indeed, OpenSSL's is a straightforward implementation of these steps. Looking at the call stack for the `BN_mod_exp_mont` break point, the function `BN_is_prime_fasttest_ex` implements iterating this test for different $b$ values to obtain prime confidence after sufficient successful trials. It carries out each trial by calling the function `witness` that performs the modular exponentiation, unfortunately calling `BN_mod_exp_mont` without setting `BN_FLG_CONSTTIME`. The algorithm continues with a classical sliding window exponentiation, potentially leaking partial information on $x$ hence $p$. Note that the sliding window code path is known to be vulnerable to cache-timing attacks and was first exploited by Percival [Per05], nevertheless this leak is not relevant for our attack.

**Insecure inversion code path.** Related to the previous code path, as the function name `BN_mod_exp_mont` suggests, the implementation uses Montgomery arithmetic for efficiency. The Montgomery setup phase occurs in `BN_MONT_CTX_set`, computing the inverse of $2^w$ modulo $p$ for $w$-bit architectures. Examining the call stack, the function calls `BN_mod_inverse` without setting `BN_FLG_CONSTTIME`, potentially leaking critical binary GCD algorithm state. However, in this case our terse analysis reveals the operands are not $\{2^w, p\}$ but $\{2^w, p \bmod 2^w\}$, implemented by copying the least significant word of $p$ to a temporary BIGNUM. While all leaks are bad, some are worse than others—this is a nominal leak on the least significant word of $p$.

**Insecure GCD code path.** The shallowest call stack is for `BN_gcd`, called directly by `rsa_builtin_keygen` (Line 154). The function computes the GCD of $e$ and $p - 1$ to ensure that $e$ is invertible $\bmod (p-1)(q-1)$. The value $p - 1$ should remain secret, hence hitting this break point represents a potential side-channel attack vector. This is the code path we target in the remainder of this paper due to its novelty compared to insecure exponentiation.

**Root cause analysis.** From these results, we deduce modular inversions in OpenSSL's RSA key generation at Steps 6 and 9 of Algorithm 1 have side-channel mitigations in place, yet GCD computations in Steps 2 and 4 lack such protection, and likewise for primality testing. The work of Acıiçmez, Gueron, and Seifert [AGS07] induced the secure path code change, yet the impact of the academic result did not fully propagate throughout the

entirety of the RSA key generation implementation. We speculate this is a result of a simplification in Acıiçmez, Gueron, and Seifert [AGS07, Sec. 2.1]: the pseudocode for key generation abstracts away the prime generation loop, and assumes a priori coprimality of $e$ with $p-1$ and $q-1$ to compute $d$ at Step 6. This allowed the authors to focus theoretical analysis on the impact of modular inversion leaks across various cryptosystems, while undoubtedly being aware of GCD and modular inversion execution flows having essentially the same branching characteristics. Alas, typical engineers are less inclined to such cryptographic subtleties—evidenced by this code path remaining vulnerable to microarchitecture side-channel attacks.

**The tool.** Subsequent to our work, Gridin et al. [Gri+19] expanded our methodology and tooling into a full-fledged Continuous Integration (CI) tool named *Triggerflow*[4]. It offers a different approach compared to static program analysis tools [DK17, Doy+15, Ant+17], and dynamic program analysis tools [Wan+17, Wic+18, Wei+18]. Rather than automated detection of security vulnerabilities and leakage quantification, our tool works in a white-box model where it complements other tools and assists developers to find undesired execution flows—such as non constant-time algorithm executions—and reports them back to the developers for further analysis. See [Gri+19] for more information about the goals, uses, and limitations of the tool.

## 3.2 Theoretical Leakage Analysis

Pereida García and Brumley [PGB17] demonstrate it is possible to recover some $Z_i$ from OpenSSL modular inversion operations (BEEA) with cache timings during ECDSA signature generations. We are left with the following open question: *Is it possible to similarly recover binary GCD algorithm state?*

Aldaya et al. [Ald+17] analyzed RSA key generation with respect to SCA of GCD-based algorithms. The analysis focuses on the modular inversion at Step 6 of Algorithm 1, exploiting the fact that BEEA inputs have very different bit-lengths. The product $(p-1)(q-1)$ has 2048 bits for modern RSA key sizes, while the other input $e$ has only 17 bits commonly.

Our vulnerability similarly exploits a large bit-length difference between inputs: the same $e$, but instead $p-1$ and $q-1$ having 1024 bits. As processing $p$ and $q$ are very similar regarding GCD computation, we use the prime $p$ to present our analysis. Furthermore, we select the *partial* bit-recovery model (see Section 2.3) because (1) we will be working with noisy LS-sequences later in our full attack, thus partial recovery reduces noise influence; (2) covered later in this section, we will utilize a factoring method that inputs incomplete $p$; and (3) we need to recover hundreds of bits so the *look-up* model is intractable.

The large bit-length difference between $p-1$ and $e$ implies that during several binary GCD iterations the condition $u \geq v$ will be true, giving the adversary partial execution flow information a priori (i.e. $X_i$='$u$' for some iterations $i$). This situation holds until $u$ (initialized to $p-1$) stores a value of roughly the same bit-length as $v$ (initialized to $e$). Therefore it divides $u$ by two roughly $\lg(N)/2 - \lg e$ times.

According to the $Z_i$ definition, at each iteration $u$ loses $Z_i$ bits. Therefore the number of iterations $t$ that should execute before $u$ has roughly the same bit-length as $v$ is the minimum $t$ that satisfies (1).

$$n = \sum_{i=1}^{t} Z_i \geq \lg(N)/2 - \lg e \tag{1}$$

Hence, the following question arises: *how many bits can be recovered in this setting?*

[4]Freely available, open source: https://gitlab.com/nisec/triggerflow

**Partial recovery.** Applying the *partial* model, we obtain a bit-recovery equation as follows. Assume the adversary obtains all $Z_i$, and $t$ is the first iteration for which $u < v$ (i.e. $X_i$='$u$' ; $0 < i < t$). The values of $u$ and $v$ just before the *sub-step* for iterations $i < t$ are the following:

$$u_1 = p - 1, \qquad v_1 = v_i = e, \qquad u_{i+1} = \frac{u_i - v_i}{2^{Z_{i+1}}}$$

The invariant $u_i - v_i \equiv 0 \bmod 2$ holds for all iterations, since both variables are odd just before the *sub-step*. Expanding for $i < t$:

$$u_t - v_t = \cfrac{\cfrac{\cfrac{\cfrac{\cfrac{p-1}{2^{Z_1}} - e}{2^{Z_2}} - e}{2^{Z_3}} - e}{\ddots}}{2^{Z_t}} - e \equiv 0 \bmod 2$$

thus solving for $p$ yields (2) for bit recovery, where $n$ from (1) is the number of recovered bits from $p$.

$$p \equiv e(2^{Z_1} + 2^{Z_1+Z_2} + \cdots + 2^n) + 1 \bmod 2^{n+1} \tag{2}$$

In summary, for RSA-2048 $n$ is roughly $1024 - 17 = 1007$ bits. However, due to Coppersmith [Cop96] an adversary only needs $\lg(N)/4 = 512$ bits of one prime to factor an RSA-2048 $N$, depicted in Figure 3. For either NIST compliant value of $e$, the number of bits recovered is far beyond the Coppersmith bound.

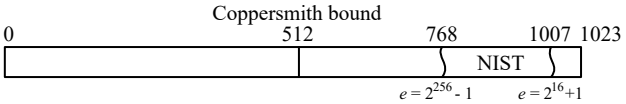


**Figure 3:** RSA-2048 bit-recovery bounds of $n$ for different $e$.

We now have our requirements for a successful attack—the adversary must obtain (at least) the first $t_c$ noise-free $Z_i$ to factor $N$:

$$n = \sum_{i=1}^{t_c} Z_i \geq \lg(N)/4$$

where $t_c$ is the minimum iteration for which $n$ reaches the Coppersmith bound.

## 3.3 A Single Trace Attack: Roadmap

Previously in this section, we uncovered three side-channel insecure code paths traversed during RSA key generation. Subsequently focusing on the `BN_gcd` code path, we then gave a theoretical analysis on the GCD algorithm as implemented in OpenSSL to describe what kind of side-channel information we can extract, and rough bounds for how much (noise-free) information we need to leverage that to recover the private key by factoring $N$. The remainder of this paper is dedicated to describing the methods, techniques and problems faced when trying to recover the necessary information from the side-channel leakage in order to achieve full private RSA key recovery from a single trace. The roadmap for our end-to-end attack is as follows: (1) We capture cache-timing traces from `BN_gcd` executions during RSA key generation, then—leveraging signal processing techniques—extract the portions corresponding to $p-1$ and $q-1$, apply digital filters and extract their corresponding (noisy) LS-sequences (Section 3.4); (2) Building upon previous work, we design and

implement an error correction algorithm for these sequences—leveraging number theoretic constraints imposed by RSA—to extract partial bits of one factor of $N$ (Section 4); (3) Said algorithm yields an ordered list of candidates for partial factors; we then derive lattice parameters for factoring with Coppersmith's method, and create lattice instances with said candidates, iteratively executing them until the result yields complete factorization of $N$ (Section 5).

**Attack setup.** Our attack setup consists of an Intel Core i5-2400 Sandy Bridge 3.10 GHz (32 nm) with 8 GB of memory running 64-bit Ubuntu 16.04 LTS "Xenial" with hardware prefetching and Turbo Boost disabled. All the cores share a 12-way 6 MB unified LLC. The system does not feature HyperThreading.

We tested our attack against OpenSSL 1.0.2k—the latest release and LTS version at the time of our experiments—with debugging symbols on the executable. We use the debugging symbols to map source code to memory addresses, allowing us to find the "hot" memory addresses for the degrading attack and probing accurately the sequence of operations mentioned previously. Note, however, debugging symbols are not a requirement for the attack as this information can be obtained through reverse engineering. We passed the `shared` configuration option to compile OpenSSL as a shared object.

## 3.4 From Timings to Sequence of Operations

The GCD algorithm implemented in OpenSSL is highly dependent on its inputs during execution, thus we use the well-known Flush+Reload technique to probe cache lines in code routines `BN_rshift1` and `BN_sub`. By probing these two routines, we are able to distinguish two branches executed by the GCD algorithm, namely right-shifts and subtractions. Unfortunately this is not enough to recover meaningful data, since we need to know the exact $Z_i$ values (i.e. number of right-shifts executed between subtractions) in order to identify bits. Due to tight loop execution during these operations, our probe misses some of the accesses.

To that end, to get better resolution we pair the Flush+Reload technique with the performance degradation attack [All+16] which targets different cache lines in the same previous routines to slow down the execution. Moreover, we apply the profiling approach [PGB17] to easily identify the best memory addresses to probe and degrade. Adapted to our strategy, this provides a good starting point to recover a sequence of operations.

**Granularity.** Due to the nature of the GCD algorithm, the granularity of the `BN_rshift1` and `BN_sub` operations captured in a trace vary throughout the execution of the algorithm. As the input values to the function are processed, their bit length decreases and this behavior is reflected in the trace. Typically, when a GCD operation begins, a single right-shift operation spans over several data points (i.e. cache-hits) in the trace, while at the end of the execution the same right-shift operation registers fewer points, sometimes even only one point. This represents a challenge later in our attack during the horizontal analysis, when we need to extract the sequence of right-shift and subtraction operations from the $p$ and $q$ traces (i.e. $Z_i$), since operations can be easily misclassified due to high data point variation for each operation within a single GCD execution.

**Traces.** The contents of a typical trace (see Figure 4, top), from high to low abundance, is roughly: (1) noise and/or `BN_mod_exp_mont` executions during primality testing, not targeted by our probes but nonetheless consuming CPU cycles; (2) short `BN_mod_inverse` executions setting up Montgomery arithmetic, with the same underlying right-shifts and subtracts as `BN_gcd` that our probes target; (3) longer `BN_gcd` executions for testing

coprimality. We are only interested in the latter, yet manually isolating the part of the trace that corresponds to the real `BN_gcd` executions for $p-1$ and $q-1$ is time consuming and not feasible at a large scale.

Unlike other side-channel scenarios where attackers have oracle access to trigger the cryptographic operation, i.e. they can start side-channel signal acquisition (e.g. launch the spy) roughly at the same time as they start their query (e.g. initiate a protocol run or call an API), our case is very different. In our case, we have no control over when key generation takes place, leaving us with extremely long traces, and the challenging task of extracting a minuscule partial trace from abundant data—looking for a needle in a haystack. We turn to signal processing-based power analysis techniques to tackle this issue.

**Templates.** Inspired by SPA and template attacks [CRR02] and related (yet less statistical) cache-based techniques [BH09, GSM15], we create a template by manually adjusting Flush+Reload parameters in our spy process, then inspecting and trimming a trace that looks visually correct, i.e. no clear preemptions of the spy or the victim process.

**Correlation and matching.** Motivated by statistical methods such as Horizontal Correlation Analysis [Cla+10], we leverage the Pearson correlation coefficient to find the best match for these templates within full traces, automating the process. Once we create our template and acquire our trace, we compute the moving Pearson correlation between the template and full trace, then extract the peaks, i.e. discrete indices that exhibit highest correlation between the template and the side-channel data. This automates the GCD identification step, allowing us to extract the sequences for $p$ and $q$.

**Horizontal analysis.** Finally, to overcome the granularity issues previously mentioned, we use a dynamic horizontal analysis approach to recover the sequence of operations executed by the GCD algorithm. Our dynamic horizontal analysis works as follows: first we take as input the processed trace which has been aligned to the first subtraction operation and trimmed to a specific length, avoiding noise as much as possible. Then, we take small chunks of the trace containing $n$ windows of subtraction and right-shift operations, recall that each subtraction is followed by at least one right-shift operation. After that, we compute the Euclidean distance between the subtraction operations in those $n$ windows. We sort the resulting distances from shortest to longest and then we consider the shortest distance as a single right-shift operation (i.e. $Z_i = 1$), thus using it as basis to determine the number of right-shift operations in the rest of windows. After calculating all the right-shift counts in those $n$ windows, we proceed to the next chunk and repeat the process. We continue until all the operations in the trace have been calculated, resulting in a tentative and noisy *LS sequence* of operations (i.e. $Z_i$ candidates).

Figure 4 illustrates our method in action using a real trace and template. The top most plot is the filtered partial trace, containing two GCD runs for $p$ and $q$—note the narrower peaks corresponding to executions of `BN_mod_inverse` during the primality test, also leaking secret information on the prime values due to yet another flag not set (see Section 3.1). The third plot is the moving Pearson correlation coefficient between the template (second plot) and the trace (first plot), with two extremely distinguishable peaks that identify the locations of the two GCD operations. The second plot aligns the template at maximum correlation for visualization purposes; and finally, the bottom plot shows a closer view of the sequence of operations performed during a single GCD operation. As seen, our technique is remarkably effective.

As mentioned previously, the accuracy of the operations in the trace decreases dramatically by the end of the GCD execution. The trace contains errors introduced by several factors such as victim preemptions, spy preemptions, as well as noise created by other

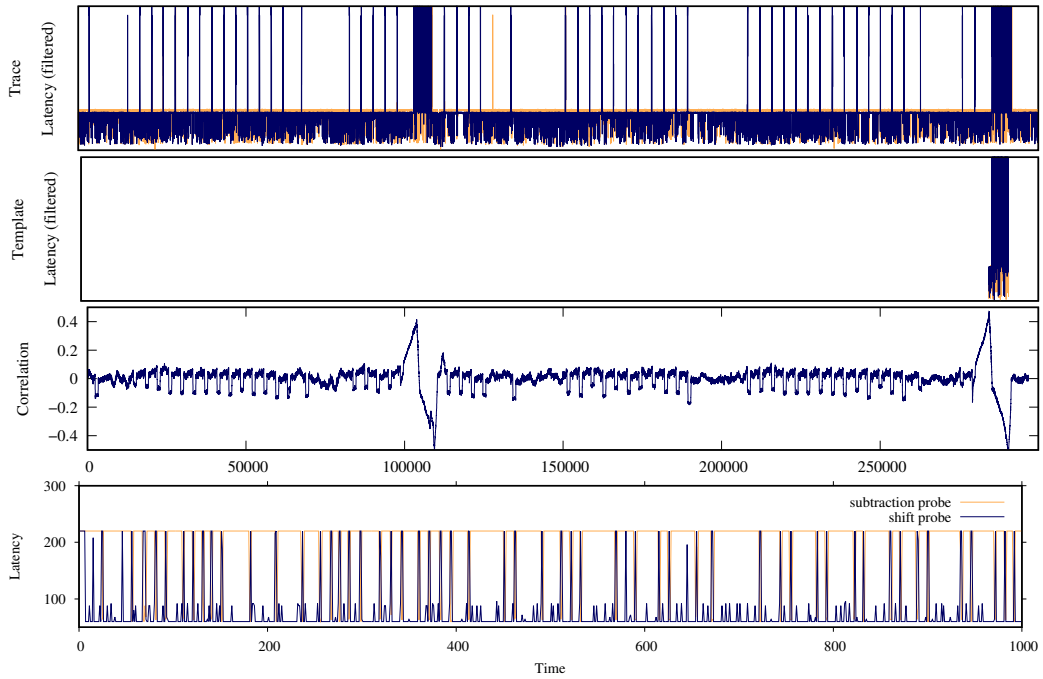


**Figure 4:** Visualization of the moving Pearson correlation in action. From top to bottom: filtered trace, aligned template trace, Pearson correlation, and raw trace (zoomed).

processes and microarchitecture components. To overcome this, Section 4 details an error correction algorithm developed to find potential correct LS sequence candidates that later are converted to bits and used as input values to perform the lattice attack explained in Section 5.

# 4 Error Correction in noisy LS Sequences

In order to design an algorithm for correcting the errors in an LS sequence, we characterize the nature of them. As discussed in Section 3.3, cache-timing attacks like Flush+Reload provide noisy data due to variances in the execution environment, interruptions, preemptions, task scheduling, etc. Therefore LS sequence extraction from the raw traces is not error free and contains errors with overwhelming probability.

After analyzing many of the traces with known inputs, we identified the following classes of errors: (1) Wrong number of 'L' symbols between two 'S' symbols (due to $Z_i$ estimation error). (2) Missing 'S' symbols (less frequent). (3) Extra 'S' symbols (much less frequent). (4) Victim preemption: observed as a small gap (i.e. window of cache misses) in the middle of operations but fixable by removing this window during trace processing. (5) Spy preemptions: observed as a *hole* in the trace exhibited by the timing information from the Flush+Reload attack. They are detectable but unfortunately operations during the preemption window are completely lost.

## 4.1 Leakage Data: Error Modeling

From the Flush+Reload attack, we obtain pairs $(Z_i^p, Z_i^q)$ for $p$ and $q$ respectively. With this information, we obtain two recovery equations according to (2), where wlog. $n_1$ and $n_2$ are greater than some $n$. Thus, we have sufficient (noisy) $Z_i$ for both primes to recover

their first $n$ bits.

$$\begin{aligned} p &\equiv ex_1 + 1 \mod 2^{n_1+1}\ ,\ x_1 = (2^{Z_1^p} + 2^{Z_1^p+Z_2^p} + \cdots + 2^{n_1}) \\ q &\equiv ex_2 + 1 \mod 2^{n_2+1}\ ,\ x_2 = (2^{Z_1^q} + 2^{Z_1^q+Z_2^q} + \cdots + 2^{n_2}) \end{aligned} \tag{3}$$

Therefore, the $Z_i$ for a prime $p$ (resp. $q$) defines set bits in the binary representation of $x_1$ (resp. $x_2$). Hence for every value of $n$ we can check if (4) holds.

$$N \equiv (ex_1 + 1)(ex_2 + 1) \mod 2^n \tag{4}$$

This relation is very similar to that employed by Heninger and Shacham [HS09] for the $m = 2$ case. In their work, errors are in the bit positions of $p$ and $q$ directly while in our case they are instead in the bit positions of a divisor of $p - 1$ and $q - 1$. Irrespective of this difference, it leads to a similar error correction algorithm constructed for correcting some errors.

Regarding existing noise models, our data might have some form of erasures due to spy preemption. However, while *erasure model* considers missing data, at the same time it requires knowing where the missing bits are and they should be distributed at random. Spy preemption fulfills the first condition but it implies a consecutive missing bits/symbols instead of random. On the other hand this model does not consider insertions and deletions.

At the same time, the bit-flip model by Henecka, May, and Meurer [HMM10] does not apply to our case as the largest error source is due to insertions and deletions of zeros between consecutive ones in the binary representation of $x_1$ and $x_2$. It is worth noting that these insertions and deletions generate some bit-flips on $p$ and $q$. However, we verified that even a small insertion/deletion rate of 0.08 implies a bit-flip rate on $p$ and $q$ of about 0.5 due to an avalanche effect. Hence, obtained $p$ and $q$ from their noisy $Z_i$ look like random data.

In this regard, the solution to bit-flip noise model by Henecka, May, and Meurer [HMM10] is tightly coupled to the Hamming distance as a metric for filtering out wrong solutions. The algorithm proposed in [HMM10] could be an option to address our noise model, but requires selecting a proper distance metric. We identified the weighted Levenshtein distance as a possible good distance candidate, as it allows assigning different weights for each operations per symbol. Then, it is possible to assign operation per symbol weights to fit a specific noise model. While this is a plausible approach, changing the distance metric inhibits us from using the theoretical and experimental results from [HMM10] to get an estimate on expected success rates. For this reason, we defer this approach to future work and implement an error correction algorithm that does not depend on an specific distance metric.

Inci et al. [Inc+16] use another interesting approach to correct errors in RSA keys. They developed an algorithm that fixes some errors in a noisy version of $dp$ and $dq$ (i.e. $m = 2$ case). Their noise model has some similarities with ours, however they considered that $dp$ is almost error-free, while in our case we cannot make this assumption. In addition, not much is said about the error rate supported by this algorithm nor its success probability under any error rate.

However, despite that *erasure* and *bit-flip* noise models do not apply directly to our scenario, we borrow the core ideas from Heninger and Shacham [HS09] and Henecka, May, and Meurer [HMM10] to build an error correction algorithm that handles the most common errors in our noise model: *insertions and deletions of zeros in the binary representation of* $x_1$ *and* $x_2$.

This relaxed noise model selection is not arbitrary. Our intuition is to use this starting approach as a building block for fixing other error sources. Also, it allows getting a first lower bound on the success rate of the attack and then scaling it (if needed) to support other errors (for example errors in 'S'). In addition, avoiding some specific error handling

like spy preemption allows using this correction algorithm in other scenarios for correcting these types of errors in other elements of `sk`.

## 4.2 Error Correction Algorithm

Our algorithm follows the Expand-and-Prune approach [HMM10] as shown in Algorithm 3. The algorithm iterates over all bits from $i \leq n$. It processes a set of candidates $\mathcal{C}_i$ at every iteration $i$, starting from a single candidate with the noisy $x_1$, $x_2$. At each iteration, each candidate resulting from the previous iteration expands to several candidates. Then, to avoid candidate space explosion, it prunes these candidates based on rules controlled by the algorithm parameters. Therefore, the configuration parameters of the **Prune** procedure manage the search space growth rate while aiming to increase the probability that the correct solution survives through iterations.

**Algorithm 3:** Error correction algorithm

**Input:** $N$, $e$, $Z_i^p$, $Z_i^q$, $n$, config parameters
**Result:** Set of $n$-bit candidates for $x_1$ and $x_2$

```
1 begin
2     x_1, x_2 ← Using Z_i^p, Z_i^q according to (3)
3     C_0 ← {(x_1, x_2)}
4     for i = 1 to n − 1 do
5         E_i ← ∅
6         foreach c ∈ C_{i−1} do
7             E_i = E_i ∪ Expand(c, i)
8         C_i ← Prune(E_i, params)
9     return C_{n−1}
```

**Expand.** To expand a given candidate $c$ at some bit $i$ (i.e. $c[i]$), consider the selected bit as a branch in a tree. If we construct a search tree, then the possibilities for the bits at any level give rise to new branches in said tree. The tree at any level $i$ contains all the partial solutions $x_1$, $x_2$ up to the $i$-th LSB. At any level $i$ there are at most six possible branching candidates that can fulfill (4), listed in Table 1.

**Table 1:** Possible branching candidates.

| Possibilities | Description |
|---|---|
| $(x_1, x_2)$ | (4) holds without changes to $x_1$ or $x_2$ |
| $(x_1 - 0, x_2)$ | Remove a zero at position $i$ from $x_1$; no changes to $x_2$ |
| $(x_1 + 0, x_2)$ | Insert a zero at position $i$ in $x_1$; no changes to $x_2$ |
| $(x_1, x_2 - 0)$ | Remove a zero at position $i$ from $x_2$; no changes to $x_1$ |
| $(x_1, x_2 + 0)$ | Insert a zero at position $i$ in $x_2$; no changes to $x_1$ |
| mult | A combination of changes in both $x_1$ and $x_2$ |

One interesting feature of this algorithm is that even if (4) holds without changing $x_1$ and $x_2$, it still tests the remaining possibilities, including errors that occur at the same index $i$ in $x_1$ and $x_2$—a situation not detected using (4).

Figure 5 illustrates the expansion and pruning procedure for three consecutive iterations of a candidate $c$ starting at bit $i$ showing the possible candidates. Here we used $\emptyset$ to represent the candidates that did not generate valid solutions. Note how in the first

expansion, four possible branching candidates fulfilled (4), however, some of them did not generate viable solutions afterwards or were pruned.

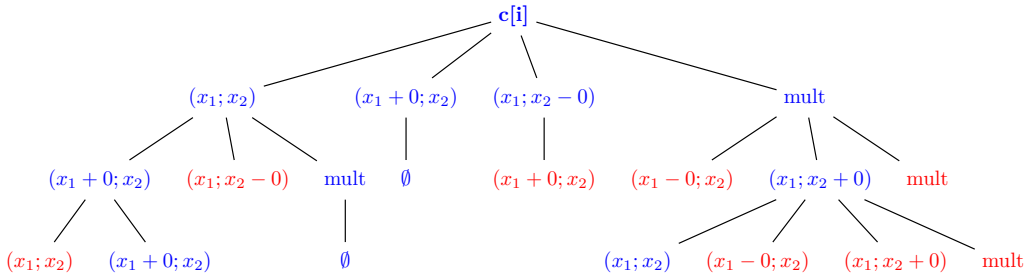


**Figure 5:** Expansion process for candidate c[i]. Pruned solutions are in red.

The candidate pool grows exponentially. We now turn to restricting the number of potential candidates (i.e. the partial solutions) at any level so that finding the correct one is possible through exhaustive search among all solutions within a certain feasible limit, examining situations that control the branching behavior of the tree.

**Prune.** The pruning process applies a set of filters to the expanded candidates $\mathcal{E}_i$. The filter spectrum is very wide and selecting the best for a given noise model is a challenging task. In this regard, contrary to [HMM10] we implemented a set of filters in a combined approach—the most novel feature of this algorithm.

Candidates in $\mathcal{E}_i$ are grouped based on the total amount of *changes* (i.e. potential errors) made in both $x_1$ and $x_2$ until iteration $i$ (see Table 1 for the list of possible changes), allowing us to sort the potential candidates based on this parameter.

The most important filters relate to the number of groups ($g$) to keep and the maximum number of candidates in each group ($G$). Denote $e_{\min}$ as the minimum number of changes made in all candidates in $\mathcal{E}_i$, $e_x$ the number of changes in $x_1$ or $x_2$, and $e_{\text{mult}}$ the number of multiple changes up to iteration $i$. We define the filters as follows: (1) *Max changes over minimum:* Keep the candidate if $e_{x_1} + e_{x_2} + e_{\text{mult}} \leq e_{\min} + g$ holds. (2) *Max candidates:* Sort each group based on $(e_{x_1} + e_{x_2} + e_{\text{mult}}, e_{\text{mult}})$, and keep the first $G$ candidates. (3) *Consecutive changes:* Discard sequences having more than a fixed number of consecutive changes, following the heuristic that higher change densities should be less probable than lower ones—therefore it should be more likely that the correct solution has a smaller change density. (4) *Max changes (hard threshold):* Candidates that exceed a maximum number of changes threshold are discarded ($e_{x_1} + e_{x_2} + e_{\text{mult}} \geq e_{\text{th}}$), helping to detect very unlikely solutions—those with an extremely high number of changes.

We selected the parameters of these filters to keep the probability of pruning the correct solution low, while at the same time keeping the computational requirements affordable for the attacker. In general terms, the adversary can profile the target environment—generating a set of known RSA keys and collecting information about the number of errors, their distribution, etc. for selecting these parameters. We followed this approach for 100 independent RSA-2048 keys and tuned these parameters for our attack environment. We analyzed the number of groups between 5 and 10 and the number of candidates in each group between 5000 and 15000. We set the number of consecutive changes filter to three and based on the observed error rates it is very unlikely that a candidate has more than 150 errors at any iteration, therefore we set the hard threshold to this value.

After this characterization, we observed that 37 traces of 100 fit our reduced noise model: only errors in the number of zeros between ones of $x_1$ and $x_2$. We recovered at least 512 bits from 30 of the test traces, therefore our correction algorithm worked for

80% of the traces it can handle (reduced noise model). However, as we used these same traces to tune the filter parameters, this value should not be taken as a measure of the success rate of our algorithm as it is biased. Section 5.1 shows the experimental results for 10K independent traces, and from this large set we extracted the estimate that our error correction algorithm recovers at least 512 bits for 73% of the traces that met our reduced noise model (see Section 5.1 for more details).

It is worth noting that these success rates and handled error rates are incompatible with other works (e.g. [HMM10]) due to different noise models with respect to previous work. However, we point out that the multiple filter approach that we follow in our algorithm could be an interesting option for addressing other noise models, where initial experiments suggest that some previous works would be improved.

**Candidates enumeration.** One important feature of our algorithm is the way it enumerates candidates for checking factors of $N$ (i.e. next stage of our end-to-end attack). As described above, each $\mathcal{C}_i$ consists of a fixed number of groups $g$ with at most $G$ possible candidates in each group. The naïve approach would be to search all possible candidates in each group until finding the solution. However, based on empirical data, we found that the real solution tends towards the first position of a group. In this case, it makes more sense to consume the candidates using a round robin approach, giving a higher priority to the highest ranked candidates in each group.

# 5 Factoring with Partial Information: Endgame

In his groundbreaking work, Coppersmith [Cop96] proposed a method to find small solutions of univariate modular equations with modulus having unknown factorization. This result finds many uses in cryptography (mainly in cryptanalysis) as several times in real-world applications an attacker has access to an oracle that gives partial information of a secret and the problem of recovering the remaining part is modeled as a univariate modular equation.

Side-channel attacks play very nicely the oracle role as they often only reveal a (minority) fraction of secret bits. Also, as in our scenario even if it is theoretically possible to fully recover the primes from side-channel traces, it is preferable to only partially recover them to reduce noise influence.

Coppersmith's result has several implications on RSA security. For excellent surveys about its impact, we refer readers to [NV10, Hin10]. One of these applications is factoring $N$ when half the bits of one prime are known—either the most or the least significant half. The lattice-based solution to this application has been extensively covered in literature [NV10, Hin10, Nem+17]. However, for the sake of completeness Appendix A contains a full description of this procedure and its parametrization in terms of how many bits of $p$ are needed to factor an RSA-2048 modulus. To summarize, we estimate that 522 bits of $p$ are sufficient to compromise RSA-2048 with high probability.

## 5.1 Results: End-to-End Attack

To consolidate the attack and validate the successfulness of our techniques and the attack overall, we used the following setup: a core executing 10K RSA key generations using OpenSSL `genpkey` command, while degrading the performance in two additional cores and finally executing the spy process on the last core, thus collecting 10K traces. Once we collected the traces, and using templates and the Pearson correlation coefficient, we extracted and aligned the GCD operations for $p-1$ and $q-1$. Out of those 10K traces, 566 traces were useless due to two main reasons: (1) key generation execution took more time than expected due to failed primality tests; or (2) spy/victim were preempted for a

long period of time; thus the spy missed capturing one or both of the GCD operations in any of those cases. We were able to perform horizontal analysis on the remaining 9434 traces to extract a tentative LS sequence of operations, i.e. $Z_i$ values, aiming to recover a minimum of 522 bits per prime value.

We then offloaded the data for these 9434 trials to a cluster for analysis, containing roughly 1500 nodes mixed between Intel E5-2680 (Sandy Bridge), E5-2670 (Sandy Bridge), and E5-2630 (Haswell) cores. In all the stages that follow, we limited per-job execution times to 4 CPU hours. Table 3 shows computation effort statistics, where a single attack run takes less than 2 CPU hours on average, making it very affordable in practice.

The LS sequences contain errors that the subsequent lattice attack cannot tolerate—error correction is required to recover sufficient bits. Our lattice attack can factor RSA-2048 knowing 522 bits of a prime, hence we configured our error correction algorithm to recover the same number of bits. Following the algorithm tuning process (see Section 4.2) we selected the set of pruning filter parameters shown in Table 2.

**Table 2:** Parameters for error correction algorithm.

| Parameter | Value |
|---|---|
| Max changes over minimum | 10 |
| Max candidates per group | 15000 |
| Consecutive changes | 3 |
| Max changes (hard threshold) | 150 |

On the cluster, we launched the algorithm for the 9434 traces that contained tentative LS sequences using Table 2 parameters. For each of the 9434 traces, we also analyzed the ground truth correct sequence to collect data about the probability of recovering a given amount of bits up to 552. Figure 6 (Left) shows the resulting survival probability curve.

This survival curve gives an idea about the error correcting algorithm behavior for our large data set. One of the most relevant results is the probability of recovering at least 522 bits: 27.89% (2632 of 9434). At the same time, the probability is quite close to that of 512 bits: 29.17%. This small difference confirms the estimation made during the lattice attack parameter optimization: correcting errors up to 522 bits is not significantly more challenging than the 512 bits case. Therefore in our setting, improving lattice attack parameters to Coppersmith's bound (512 bits) will not significantly increase the error correction success rate.

Of interest, Figure 6 (Left) shows an abrupt probability drop at the start of the curve. We analyzed it closely and confirmed that roughly 30% of the traces have a spy preemption just at the beginning, resulting in incomplete traces. It seems there is a bias in our environment that increases the probability of spy preemption at the start of a GCD execution, yet not in the middle. We are investigating the reasons behind this bias, but the fact that the Figure 6 (Left) curve does not contain another abrupt drop confirms our bias hypothesis.

After this analysis, our error correction algorithm was able to recover 522 bits for 2632 traces. One interesting metric is the number of errors considering both LS sequences (i.e. $p$ and $q$) that it handled per key, and Figure 6 (Middle) shows the boxplots of the aggregate number of errors in both LS sequences for recovering various bit quantities. The data suggest the number of errors successfully handled is diverse—for example, recovering 522 bits (rightmost boxplot), this metric ranges from 24 to 108. It implies that our error correction algorithm with Table 2 parameters recovered 522 bits in 2632 traces with error rates that range from $24/(2 \cdot 522) = 0.02$ to $108/(2 \cdot 522) = 0.10$, since we require 522 bits of each prime (i.e. $p$ and $q$).

Of these 9434 instances, we successfully recovered 2285 private keys after 12875 lattice trials; Figure 6 (Right) "Computed" depicts these data points. This represents just over

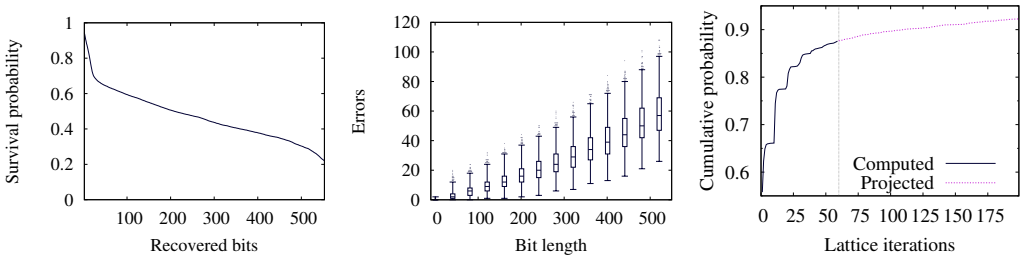


**Figure 6:** Left: bit recovery survival plot. Middle: number of errors successfully handled by our error correction algorithm. Right: Lattice iterations for successful instances.

**Table 3:** Cluster computation effort in CPU minutes across phases and CPUs.

| | Error correction phase | | | Lattice phase | | |
|---|---|---|---|---|---|---|
| CPU | Median | Mean | Dev. | Median | Mean | Dev. |
| E5-2670 | 103.0 | 103.7 | 41.2 | 6.2 | 6.7 | 1.7 |
| E5-2680 | 92.3 | 90.4 | 36.1 | 4.3 | 4.3 | 0.3 |
| E5-2630 | 100.0 | 95.0 | 40.1 | 6.8 | 7.3 | 1.8 |

45 days of CPU time. The remaining 7149 instances break down as follows, which we analyzed using the ground truth private keys. For 347 instances, the partial prime factor remained amongst the candidates, yet had a poor ranking, hence the number of lattice iterations needed exceeding our fixed allowed time (4 h); Figure 6 (Right) "Projected" depicts these data points. In these cases, we verified the lattice output at that future iteration indeed yields the intended factor. The remaining 6802 instances failed to retain the correct candidate. To find a solution, the error correction algorithm with round-robin enumeration achieved an impressive median of one lattice iteration for successful instances.

**End-to-end attack summary.** From a large data set of 9434 independent traces, our end-to-end attack achieves a success rate of 27.89%. The error correction algorithm was able to fix/recover 522 bits for 2632 traces. At the same time, the lattice attack succeeded for all these 2632 traces, showing the robustness of our lattice attack parameters for 522 bits (Table 4).

# 6 Conclusion

In this work, we proposed a methodology to analyze cryptographic software for traversal of known side-channel insecure code paths. Applying our methodology to RSA key generation in OpenSSL uncovered three new vulnerabilities, one of which we designed an end-to-end cache-timing attack around, leading to key recovery with good probability and modest computational effort. The attack chain consisted of (1) gathering timings with a combination of Flush+Reload and performance degradation; (2) locating the trace segments of interest (two specific GCD executions) within abundant data; (3) transforming these traces into noisy LS-sequences representing GCD algorithm state; (4) executing our error correction algorithm, resulting in a ranked list of partial prime factor candidates; (5) formulating lattice problems for these candidates that recover the unknown portion; (6) testing if the result yields a prime factor of the RSA modulus $N$, hence the private key. Executing 10K trials and moving the analysis to a cluster, we achieved roughly a 27% success rate for full key recovery. We close with lessons learned from our work.

**Lesson 1: Secure by default.** Similar to two recent works [PGBY16, PGB17], two of the vulnerabilities our methodology uncovered are due to insecure default behavior—failure to set a particular flag that, by early exit, diverts the code through algorithms with SCA mitigations. Had the logic been inverted, taking the secure paths by default would have prevented these vulnerabilities. For OpenSSL, these new vulnerabilities continue an unfortunate trend of insecure by default failures that went undetected during unit testing.

**Lesson 2: Knowledge transfer.** Our end-to-end attack exploits only one of the three vulnerabilities our methodology uncovered. The function we targeted is oblivious to the constant-time flag, hence having it set or clear has no effect on our attack. Our root cause analysis (Section 3.1) suggests that the mitigations mainlined as a result of pioneering academic work [AGS07] failed to consider RSA key generation as a whole, and the similarities between GCD computation (which we exploited) and modular inversion with respect to branching behavior went unnoticed when these mitigations were independently developed. This disconnect demonstrates the critical importance of engineers working side-by-side with cryptographers to ensure that academic results reach their intended impact on real-world products.

**Affected versions and responsible disclosure.** The issues presented in this paper affected OpenSSL versions 1.1.0-1.1.0h and 1.0.2-1.0.2o. Following responsible disclosure procedures, we reported these issues to OpenSSL and provided fixes, subsequently merged into the 1.0.2 and 1.1.0 branches after the embargo lifted. OpenSSL 1.1.1 did not exist at that time, thus it was never impacted. OpenSSL assigned CVE-2018-0737 based on our work.

**Acknowledgments.** We would like to thank to Matúš Nemec for sharing his results. We thank Tampere Center for Scientific Computing (TCSC) for generously granting us access to computing cluster resources. Supported in part by Academy of Finland grant 303814. The second author was supported in part by a Nokia Foundation Scholarship and by the Pekka Ahonen Fund through the Industrial Research Fund of Tampere University of Technology. This article is based in part upon work from COST Action IC1403 CRYPTACUS, supported by COST (European Cooperation in Science and Technology). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 804476).

## References

[ACSS17] Alejandro Cabrera Aldaya, Alejandro J. Cabrera Sarmiento, and Santiago Sánchez-Solano. "SPA vulnerabilities of the binary extended Euclidean algorithm". In: *J. Cryptographic Engineering* 7.4 (2017), pp. 273–285. DOI: 10.1007/s13389-016-0135-4. URL: https://doi.org/10.1007/s13389-016-0135-4.

[AGS07] Onur Acıiçmez, Shay Gueron, and Jean-Pierre Seifert. "New Branch Prediction Vulnerabilities in OpenSSL and Necessary Software Countermeasures". In: *Cryptography and Coding, 11th IMA International Conference, Cirencester, UK, December 18-20, 2007, Proceedings*. Ed. by Steven D. Galbraith. Vol. 4887. Lecture Notes in Computer Science. Springer, 2007, pp. 185–203. DOI: 10.1007/978-3-540-77272-9_12. URL: https://doi.org/10.1007/978-3-540-77272-9_12.

[Ald+17] Alejandro Cabrera Aldaya et al. “Side-channel analysis of the modular inversion step in the RSA key generation algorithm”. In: *I. J. Circuit Theory and Applications* 45.2 (2017), pp. 199–213. DOI: 10.1002/cta.2283. URL: https://doi.org/10.1002/cta.2283.

[All+16] Thomas Allan et al. “Amplifying side channels through performance degradation”. In: *Proceedings of the 32nd Annual Conference on Computer Security Applications, ACSAC 2016, Los Angeles, CA, USA, December 5-9, 2016*. Ed. by Stephen Schwab, William K. Robertson, and Davide Balzarotti. ACM, 2016, pp. 422–435. DOI: 10.1145/2991079.2991084. URL: http://doi.acm.org/10.1145/2991079.2991084.

[Ant+17] Timos Antonopoulos et al. “Decomposition instead of self-composition for proving the absence of timing channels”. In: *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2017, Barcelona, Spain, June 18-23, 2017*. Ed. by Albert Cohen and Martin T. Vechev. ACM, 2017, pp. 362–375. DOI: 10.1145/3062341.3062378. URL: https://doi.org/10.1145/3062341.3062378.

[AT07] Sarang Aravamuthan and Viswanatha Rao Thumparthy. “A Parallelization of ECDSA Resistant to Simple Power Analysis Attacks”. In: *Proceedings of the Second International Conference on COMmunication System softWAre and MiddlewaRE (COMSWARE 2007), January 7-12, 2007, Bangalore, India*. Ed. by Sanjoy Paul, Henning Schulzrinne, and G. Venkatesh. IEEE, 2007. DOI: 10.1109/COMSWA.2007.382592. URL: https://doi.org/10.1109/COMSWA.2007.382592.

[Bau+14] Aurélie Bauer et al. “Side-Channel Attack against RSA Key Generation Algorithms”. In: *Cryptographic Hardware and Embedded Systems - CHES 2014 - 16th International Workshop, Busan, South Korea, September 23-26, 2014. Proceedings*. Ed. by Lejla Batina and Matthew Robshaw. Vol. 8731. Lecture Notes in Computer Science. Springer, 2014, pp. 223–241. DOI: 10.1007/978-3-662-44709-3_13. URL: https://doi.org/10.1007/978-3-662-44709-3_13.

[Ber+13] Daniel J. Bernstein et al. “Factoring RSA Keys from Certified Smart Cards: Coppersmith in the Wild”. In: *Advances in Cryptology - ASIACRYPT 2013 - 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, December 1-5, 2013, Proceedings, Part II*. Ed. by Kazue Sako and Palash Sarkar. Vol. 8270. Lecture Notes in Computer Science. Springer, 2013, pp. 341–360. DOI: 10.1007/978-3-642-42045-0_18. URL: https://doi.org/10.1007/978-3-642-42045-0_18.

[Ber+17] Daniel J. Bernstein et al. “Sliding Right into Disaster: Left-to-Right Sliding Windows Leak”. In: *Cryptographic Hardware and Embedded Systems - CHES 2017 - 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*. Ed. by Wieland Fischer and Naofumi Homma. Vol. 10529. Lecture Notes in Computer Science. Springer, 2017, pp. 555–576. DOI: 10.1007/978-3-319-66787-4_27. URL: https://doi.org/10.1007/978-3-319-66787-4_27.

[BH09] Billy Bob Brumley and Risto M. Hakala. “Cache-Timing Template Attacks”. In: *Advances in Cryptology - ASIACRYPT 2009, 15th International Conference on the Theory and Application of Cryptology and Information Security, Tokyo, Japan, December 6-10, 2009. Proceedings*. Ed. by Mitsuru Matsui. Vol. 5912. Lecture Notes in Computer Science. Springer, 2009, pp. 667–684. DOI: 10.1007/978-3-642-10366-7_39. URL: https://doi.org/10.1007/978-3-642-10366-7_39.

[BM03] Johannes Blömer and Alexander May. "New Partial Key Exposure Attacks on RSA". In: *Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings*. Ed. by Dan Boneh. Vol. 2729. Lecture Notes in Computer Science. Springer, 2003, pp. 27–43. DOI: 10.1007/978-3-540-45146-4_2. URL: https://doi.org/10.1007/978-3-540-45146-4_2.

[Bri+06] Ernie Brickell et al. "Software mitigations to hedge AES against cache-based software side channel vulnerabilities". In: *IACR Cryptology ePrint Archive* 2006.52 (2006). URL: http://eprint.iacr.org/2006/052.

[Cla+10] Christophe Clavier et al. "Horizontal Correlation Analysis on Exponentiation". In: *Information and Communications Security - 12th International Conference, ICICS 2010, Barcelona, Spain, December 15-17, 2010. Proceedings*. Ed. by Miguel Soriano, Sihan Qing, and Javier López. Vol. 6476. Lecture Notes in Computer Science. Springer, 2010, pp. 46–61. DOI: 10.1007/978-3-642-17650-0_5. URL: https://doi.org/10.1007/978-3-642-17650-0_5.

[Cop96] Don Coppersmith. "Finding a Small Root of a Univariate Modular Equation". In: *Advances in Cryptology - EUROCRYPT '96, International Conference on the Theory and Application of Cryptographic Techniques, Saragossa, Spain, May 12-16, 1996, Proceeding*. Ed. by Ueli M. Maurer. Vol. 1070. Lecture Notes in Computer Science. Springer, 1996, pp. 155–165. DOI: 10.1007/3-540-68339-9_14. URL: https://doi.org/10.1007/3-540-68339-9_14.

[Cop97] Don Coppersmith. "Small Solutions to Polynomial Equations, and Low Exponent RSA Vulnerabilities". In: *J. Cryptology* 10.4 (1997), pp. 233–260. DOI: 10.1007/s001459900030. URL: https://doi.org/10.1007/s001459900030.

[CRR02] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. "Template Attacks". In: *Cryptographic Hardware and Embedded Systems - CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002, Revised Papers*. Ed. by Burton S. Kaliski Jr., Çetin Kaya Koç, and Christof Paar. Vol. 2523. Lecture Notes in Computer Science. Springer, 2002, pp. 13–28. DOI: 10.1007/3-540-36400-5_3. URL: https://doi.org/10.1007/3-540-36400-5_3.

[DK17] Goran Doychev and Boris Köpf. "Rigorous analysis of software countermeasures against cache attacks". In: *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2017, Barcelona, Spain, June 18-23, 2017*. Ed. by Albert Cohen and Martin T. Vechev. ACM, 2017, pp. 406–421. DOI: 10.1145/3062341.3062388. URL: https://doi.org/10.1145/3062341.3062388.

[Doy+15] Goran Doychev et al. "CacheAudit: A Tool for the Static Analysis of Cache Side Channels". In: *ACM Trans. Inf. Syst. Secur.* 18.1 (2015), 4:1–4:32. DOI: 10.1145/2756550. URL: https://doi.org/10.1145/2756550.

[FGS09] Thomas Finke, Max Gebhardt, and Werner Schindler. "A New Side-Channel Attack on RSA Prime Generation". In: *Cryptographic Hardware and Embedded Systems - CHES 2009, 11th International Workshop, Lausanne, Switzerland, September 6-9, 2009, Proceedings*. Ed. by Christophe Clavier and Kris Gaj. Vol. 5747. Lecture Notes in Computer Science. Springer, 2009, pp. 141–155. DOI: 10.1007/978-3-642-04138-9_11. URL: https://doi.org/10.1007/978-3-642-04138-9_11.

[Fip] *Digital Signature Standard (DSS)*. FIPS PUB 186-4. National Institute of Standards and Technology, 2013. DOI: 10.6028/NIST.FIPS.186-4. URL: https://doi.org/10.6028/NIST.FIPS.186-4.

[Gri+19] Iaroslav Gridin et al. “Triggerflow: Regression Testing by Advanced Execution Path Inspection”. In: *Detection of Intrusions and Malware, and Vulnerability Assessment - 16th International Conference, DIMVA 2019, Gothenburg, Sweden, June 19-20, 2019, Proceedings*. Ed. by Roberto Perdisci et al. Vol. 11543. Lecture Notes in Computer Science. Springer, 2019, pp. 330–350. DOI: 10.1007/978-3-030-22038-9_16. URL: https://doi.org/10.1007/978-3-030-22038-9_16.

[GSM15] Daniel Gruss, Raphael Spreitzer, and Stefan Mangard. “Cache Template Attacks: Automating Attacks on Inclusive Last-Level Caches”. In: *24th USENIX Security Symposium, USENIX Security 15, Washington, D.C., USA, August 12-14, 2015*. Ed. by Jaeyeon Jung and Thorsten Holz. USENIX Association, 2015, pp. 897–912. URL: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/gruss.

[GVY17] Daniel Genkin, Luke Valenta, and Yuval Yarom. “May the Fourth Be With You: A Microarchitectural Side Channel Attack on Several Real-World Applications of Curve25519”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. Ed. by Bhavani M. Thuraisingham et al. ACM, 2017, pp. 845–858. DOI: 10.1145/3133956.3134029. URL: http://doi.acm.org/10.1145/3133956.3134029.

[Hen+12] Nadia Heninger et al. “Mining Your Ps and Qs: Detection of Widespread Weak Keys in Network Devices”. In: *Proceedings of the 21th USENIX Security Symposium, Bellevue, WA, USA, August 8-10, 2012*. Ed. by Tadayoshi Kohno. USENIX Association, 2012, pp. 205–220. URL: https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/heninger.

[Hin10] M. Jason Hinek. *Cryptanalysis of RSA and its variants*. Chapman & Hall/CRC Cryptography and Network Security. CRC Press, 2010. ISBN: 978-1-4200-7518-2. DOI: 10.1201/9781420075199. URL: https://doi.org/10.1201/9781420075199.

[HMM10] Wilko Henecka, Alexander May, and Alexander Meurer. “Correcting Errors in RSA Private Keys”. In: *Advances in Cryptology - CRYPTO 2010, 30th Annual Cryptology Conference, Santa Barbara, CA, USA, August 15-19, 2010. Proceedings*. Ed. by Tal Rabin. Vol. 6223. Lecture Notes in Computer Science. Springer, 2010, pp. 351–369. DOI: 10.1007/978-3-642-14623-7_19. URL: https://doi.org/10.1007/978-3-642-14623-7_19.

[How97] Nick Howgrave-Graham. “Finding Small Roots of Univariate Modular Equations Revisited”. In: *Cryptography and Coding, 6th IMA International Conference, Cirencester, UK, December 17-19, 1997, Proceedings*. Ed. by Michael Darnell. Vol. 1355. Lecture Notes in Computer Science. Springer, 1997, pp. 131–142. DOI: 10.1007/BFb0024458. URL: https://doi.org/10.1007/BFb0024458.

[HS09] Nadia Heninger and Hovav Shacham. “Reconstructing RSA Private Keys from Random Key Bits”. In: *Advances in Cryptology - CRYPTO 2009, 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009. Proceedings*. Ed. by Shai Halevi. Vol. 5677. Lecture Notes in Computer Science. Springer, 2009, pp. 1–17. DOI: 10.1007/978-3-642-03356-8_1. URL: https://doi.org/10.1007/978-3-642-03356-8_1.

[IES16] Gorka Irazoqui, Thomas Eisenbarth, and Berk Sunar. "Cross Processor Cache Attacks". In: *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2016, Xi'an, China, May 30 - June 3, 2016*. Ed. by Xiaofeng Chen, XiaoFeng Wang, and Xinyi Huang. ACM, 2016, pp. 353–364. DOI: 10.1145/2897845.2897867. URL: http://doi.acm.org/10.1145/2897845.2897867.

[Inc+16] Mehmet Sinan Inci et al. "Cache Attacks Enable Bulk Key Recovery on the Cloud". In: *Cryptographic Hardware and Embedded Systems - CHES 2016 - 18th International Conference, Santa Barbara, CA, USA, August 17-19, 2016, Proceedings*. Ed. by Benedikt Gierlichs and Axel Y. Poschmann. Vol. 9813. Lecture Notes in Computer Science. Springer, 2016, pp. 368–388. DOI: 10.1007/978-3-662-53140-2_18. URL: https://doi.org/10.1007/978-3-662-53140-2_18.

[Kle+10] Thorsten Kleinjung et al. "Factorization of a 768-Bit RSA Modulus". In: *Advances in Cryptology - CRYPTO 2010, 30th Annual Cryptology Conference, Santa Barbara, CA, USA, August 15-19, 2010. Proceedings*. Ed. by Tal Rabin. Vol. 6223. Lecture Notes in Computer Science. Springer, 2010, pp. 333–350. DOI: 10.1007/978-3-642-14623-7_18. URL: https://doi.org/10.1007/978-3-642-14623-7_18.

[Koc+19] Paul Kocher et al. "Spectre Attacks: Exploiting Speculative Execution". In: *2019 IEEE Symposium on Security and Privacy, SP 2019, Proceedings, 20-29 May 2019, San Francisco, California, USA*. IEEE, 2019, pp. 19–37. DOI: 10.1109/SP.2019.00002. URL: https://doi.org/10.1109/SP.2019.00002.

[Kun15] Noboru Kunihiro. "An Improved Attack for Recovering Noisy RSA Secret Keys and Its Countermeasure". In: *Provable Security - 9th International Conference, ProvSec 2015, Kanazawa, Japan, November 24-26, 2015, Proceedings*. Ed. by Man Ho Au and Atsuko Miyaji. Vol. 9451. Lecture Notes in Computer Science. Springer, 2015, pp. 61–81. DOI: 10.1007/978-3-319-26059-4_4. URL: https://doi.org/10.1007/978-3-319-26059-4_4.

[Kun18] Noboru Kunihiro. "Mathematical Approach for Recovering Secret Key from Its Noisy Version". In: *Mathematical Modelling for Next-Generation Cryptography*. Vol. 29. Math. Ind. (Tokyo). Springer, Singapore, 2018, pp. 199–217. DOI: 10.1007/978-981-10-5065-7. URL: https://doi.org/10.1007/978-981-10-5065-7.

[Len+12] Arjen K. Lenstra et al. "Public Keys". In: *Advances in Cryptology - CRYPTO 2012 - 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*. Ed. by Reihaneh Safavi-Naini and Ran Canetti. Vol. 7417. Lecture Notes in Computer Science. Springer, 2012, pp. 626–642. DOI: 10.1007/978-3-642-32009-5_37. URL: https://doi.org/10.1007/978-3-642-32009-5_37.

[Lip+18] Moritz Lipp et al. "Meltdown: Reading Kernel Memory from User Space". In: *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*. Ed. by William Enck and Adrienne Porter Felt. USENIX Association, 2018, pp. 973–990. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/lipp.

[LLL82] A. K. Lenstra, H. W. Lenstra Jr., and L. Lovász. "Factoring polynomials with rational coefficients". In: *Math. Ann.* 261.4 (1982), pp. 515–534. ISSN: 0025-5831. DOI: 10.1007/BF01457454. URL: https://doi.org/10.1007/BF01457454.

[Nem+17] Matús Nemec et al. “The Return of Coppersmith’s Attack: Practical Factorization of Widely Used RSA Moduli”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. Ed. by Bhavani M. Thuraisingham et al. ACM, 2017, pp. 1631–1648. DOI: 10.1145/3133956.3133969. URL: http://doi.acm.org/10.1145/3133956.3133969.

[NV10] Phong Q. Nguyen and Brigitte Vallée, eds. *The LLL Algorithm - Survey and Applications*. Information Security and Cryptography. Springer, 2010. ISBN: 978-3-642-02294-4. DOI: 10.1007/978-3-642-02295-1. URL: https://doi.org/10.1007/978-3-642-02295-1.

[Per05] Colin Percival. “Cache Missing for Fun and Profit”. In: *BSDCan 2005, Ottawa, Canada, May 13-14, 2005, Proceedings*. 2005. URL: http://www.daemonology.net/papers/cachemissing.pdf.

[PGB17] Cesar Pereida García and Billy Bob Brumley. “Constant-Time Callees with Variable-Time Callers”. In: *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017*. Ed. by Engin Kirda and Thomas Ristenpart. USENIX Association, 2017, pp. 83–98. ISBN: 978-1-931971-40-9. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/garcia.

[PGBY16] Cesar Pereida García, Billy Bob Brumley, and Yuval Yarom. ““Make Sure DSA Signing Exponentiations Really are Constant-Time””. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*. Ed. by Edgar R. Weippl et al. ACM, 2016, pp. 1639–1650. DOI: 10.1145/2976749.2978420. URL: http://doi.acm.org/10.1145/2976749.2978420.

[Pol74] J. M. Pollard. “Theorems on factorization and primality testing”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 76.3 (1974), pp. 521–528. DOI: 10.1017/S0305004100049252.

[Pol75] J. M. Pollard. “A Monte Carlo method for factorization”. In: *BIT Numerical Mathematics* 15.3 (1975), pp. 331–334. ISSN: 1572-9125. DOI: 10.1007/BF01933667. URL: https://doi.org/10.1007/BF01933667.

[PPS12] Kenneth G. Paterson, Antigoni Polychroniadou, and Dale L. Sibborn. “A Coding-Theoretic Approach to Recovering Noisy RSA Keys”. In: *Advances in Cryptology - ASIACRYPT 2012 - 18th International Conference on the Theory and Application of Cryptology and Information Security, Beijing, China, December 2-6, 2012. Proceedings*. Ed. by Xiaoyun Wang and Kazue Sako. Vol. 7658. Lecture Notes in Computer Science. Springer, 2012, pp. 386–403. DOI: 10.1007/978-3-642-34961-4_24. URL: https://doi.org/10.1007/978-3-642-34961-4_24.

[Rab80] Michael O. Rabin. “Probabilistic algorithm for testing primality”. In: *J. Number Theory* 12.1 (1980), pp. 128–138. ISSN: 0022-314X. DOI: 10.1016/0022-314X(80)90084-0. URL: https://doi.org/10.1016/0022-314x(80)90084-0.

[RSA78] R. L. Rivest, A. Shamir, and L. Adleman. “A method for obtaining digital signatures and public-key cryptosystems”. In: *Comm. ACM* 21.2 (1978), pp. 120–126. ISSN: 0001-0782. DOI: 10.1145/359340.359342. URL: https://doi.org/10.1145/359340.359342.

[Sch+17] Michael Schwarz et al. “Malware Guard Extension: Using SGX to Conceal Cache Attacks”. In: *Detection of Intrusions and Malware, and Vulnerability Assessment - 14th International Conference, DIMVA 2017, Bonn, Germany, July 6-7, 2017, Proceedings*. Ed. by Michalis Polychronakis and Michael Meier. Vol. 10327. Lecture Notes in Computer Science. Springer, 2017, pp. 3–24. DOI: 10.1007/978-3-319-60876-1_1. URL: https://doi.org/10.1007/978-3-319-60876-1_1.

[Ste67] Josef Stein. “Computational problems associated with Racah algebra”. In: *Journal of Computational Physics* 1.3 (1967), pp. 397–405. ISSN: 00219991. DOI: 10.1016/0021-9991(67)90047-2. URL: https://doi.org/10.1016/0021-9991(67)90047-2.

[VEW12] Camille Vuillaume, Takashi Endo, and Paul Wooderson. “RSA Key Generation: New Attacks”. In: *Constructive Side-Channel Analysis and Secure Design - Third International Workshop, COSADE 2012, Darmstadt, Germany, May 3-4, 2012. Proceedings*. Ed. by Werner Schindler and Sorin A. Huss. Vol. 7275. Lecture Notes in Computer Science. Springer, 2012, pp. 105–119. DOI: 10.1007/978-3-642-29912-4_9. URL: https://doi.org/10.1007/978-3-642-29912-4_9.

[Wan+17] Shuai Wang et al. “CacheD: Identifying Cache-Based Timing Channels in Production Software”. In: *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017*. Ed. by Engin Kirda and Thomas Ristenpart. USENIX Association, 2017, pp. 235–252. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-shuai.

[Wei+18] Samuel Weiser et al. “DATA - Differential Address Trace Analysis: Finding Address-based Side-Channels in Binaries”. In: *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*. Ed. by William Enck and Adrienne Porter Felt. USENIX Association, 2018, pp. 603–620. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/weiser.

[Wic+18] Jan Wichelmann et al. “MicroWalk: A Framework for Finding Side Channels in Binaries”. In: *Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC 2018, San Juan, PR, USA, December 03-07, 2018*. ACM, 2018, pp. 161–173. DOI: 10.1145/3274694.3274741. URL: https://doi.org/10.1145/3274694.3274741.

[Wie90] Michael J. Wiener. “Cryptanalysis of short RSA secret exponents”. In: *IEEE Trans. Information Theory* 36.3 (1990), pp. 553–558. DOI: 10.1109/18.54902. URL: https://doi.org/10.1109/18.54902.

[WSB18] Samuel Weiser, Raphael Spreitzer, and Lukas Bodner. “Single Trace Attack Against RSA Key Generation in Intel SGX SSL”. In: *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018, Incheon, Republic of Korea, June 04-08, 2018*. Ed. by Jong Kim et al. ACM, 2018, pp. 575–586. DOI: 10.1145/3196494.3196524. URL: http://doi.acm.org/10.1145/3196494.3196524.

[XCP15] Yuanzhong Xu, Weidong Cui, and Marcus Peinado. “Controlled-Channel Attacks: Deterministic Side Channels for Untrusted Operating Systems”. In: *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*. IEEE Computer Society, 2015, pp. 640–656. DOI: 10.1109/SP.2015.45. URL: https://doi.org/10.1109/SP.2015.45.

[YF14] Yuval Yarom and Katrina Falkner. "FLUSH+RELOAD: A High Resolution, Low Noise, L3 Cache Side-Channel Attack". In: *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014*. USENIX Association, 2014, pp. 719–732. ISBN: 978-1-931971-15-7. URL: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/yarom.

[YGH16] Yuval Yarom, Daniel Genkin, and Nadia Heninger. "CacheBleed: A Timing Attack on OpenSSL Constant Time RSA". In: *Cryptographic Hardware and Embedded Systems - CHES 2016 - 18th International Conference, Santa Barbara, CA, USA, August 17-19, 2016, Proceedings*. Ed. by Benedikt Gierlichs and Axel Y. Poschmann. Vol. 9813. Lecture Notes in Computer Science. Springer, 2016, pp. 346–367. DOI: 10.1007/978-3-662-53140-2_17. URL: https://doi.org/10.1007/978-3-662-53140-2_17.

# A Lattice Construction and Parametrization

Coppersmith's method reduces the modular equation problem i.e. $f(x) = 0 \bmod p$ to an equation over the integers i.e. $g(x) = 0$ with roots easily found by algorithms like Berlekamp-Zassenhaus. This transformation employs lattice reduction algorithms such as LLL [LLL82] assuming the original modular equation root is small [Cop96].

Following Coppersmith's approach, Howgrave-Graham [How97] revisited the lattice construction and proposed a new method to build a lattice that allows obtaining a $g(x) = 0$ from the original modular equation $f(x) = 0 \bmod p$. Howgrave-Graham lattice construction is often preferred due to its simplicity and numerous practical advantages [How97, NV10].

**Factoring $N$ knowing LSBs of $p$.** Assume we know wlog. the $n$ LSBs of a prime $p$ that is a factor of $N$, i.e. $p$ is expressed as

$$p = \tilde{p}2^n + p_0$$

where $p_0$ is the known portion and $\tilde{p}$ the only unknown. Hence $\tilde{p}$ is a *small* root of the polynomial

$$f(x) = x2^n + p_0 \mod p \quad |\tilde{p}| \leq X$$

and in this case, *small* meaning that $\tilde{p}$ is bound by some known constant $X$. Coppersmith approach requires $f(x)$ to be monic. To achieve that, define $b = 2^{-n} \bmod N$, where $b2^n = 1 + kN$ for some integer $k$, then express $f(x)$ as follows.

$$f(x) = x + bp_0 \mod p \quad |\tilde{p}| \leq X \tag{5}$$

Coppersmith-Howgrave-Graham approach aims to solve (5) by reducing this univariate modular equation to an equation over the integers, visualized below.

$$\underbrace{f(x_0) = 0 \bmod p \Rightarrow f_i(x_0) = 0 \bmod p^m \Rightarrow B \overset{\text{LLL}}{\Rightarrow} g(x_0) = 0}_{\text{Coppersmith-Howgrave-Graham}}$$

From the monic polynomial $f(x)$, build a set of $d = m + t$ polynomials $f_i(x)$ over $p^m$ according to the Howgrave-Graham approach, such that these $f_i(x)$ have the same root $x_0 = \tilde{p}$ modulo $p^m$ as $f(x)$ modulo $p$ [How97]. Said polynomials are as follows.

$$\begin{aligned} f_i(x) &= N^i f^{m-i}(x) && i = 0, 1, \cdots, m-1 \\ f_{m+i}(x) &= x^i f^m(x) && i = 0, 1, \cdots, t-1 \end{aligned}$$

The next step builds a lattice $B$ from the $f_i(x)$ for $0 \leq i < d$. Following Howgrave-Graham [How97] the basis vectors of $B$ are the coefficient vectors of $f_i(xX)$. Then, lattice-reduced $B$ should yield a $g(x)$ over the integers if the Coppersmith-Howgrave-Graham conditions are respected or heuristically relaxed—we expand later. The small root bound $X$ defines these conditions and the lattice dimension $d = m + t$, therefore we select them such that we can factor $N$ knowing $n$ bits of $p$.

## A.1 Lattice Parametrization

Three parameters control the effectiveness and efficiency of this method: $X$, $m$ and $t$, where the latter two control the lattice dimension ($d = m + t$) hence the amount of information in it. This dimension dictates the running time of LLL, therefore the goal is to minimize $m$ and $t$ considering that the attacker should run it for each candidate resulting from the error correction phase.

To validate the Coppersmith-Howgrave-Graham approach, we used a public SageMath implementation[5]. The objective of this validation is to obtain—for $n$, a given number of LSBs—which parameter set ($X$, $m$, $t$) yields the right solution with very high probability while minimizing the runtime as much as possible. This approach is very similar to that of Nemec et al. [Nem+17], where the authors fixed the bound $X$ and optimize $m$ and $t$—however we also tweak $X$ to get some runtime improvements.

One of the main tasks for using Coppersmith-Howgrave-Graham method is selecting the bound $X$ of the unknown root. Recalling Section 2.1, $N$ of an RSA-2048 key has exactly 2048 bits by forcing the two MSBs of $p$ and $q$ to be set, i.e. they have exactly 1024 bits. Hence for RSA-2048, the inequality (6) holds, where the ordering of $p$ and $q$ is arbitrary.

$$q < \sqrt{N} < p < 2^{1024} < 2\sqrt{N} < N \tag{6}$$

Considering that we know the $n$ LSBs of $p$, we divide (6) by $2^n$ to obtain bounds for $\tilde{p}$.

$$\frac{q}{2^n} < \frac{\sqrt{N}}{2^n} < \tilde{p} < \frac{2^{1024}}{2^n} < \frac{2\sqrt{N}}{2^n} < \frac{N}{2^n}$$

This results in the bound $X = \frac{2\sqrt{N}}{2^n}$ that should work for both primes. However, in practice the Coppersmith conditions are slightly pessimistic, hence in our analysis we also consider $X = \frac{\sqrt{N}}{2^n}$.

**Optimizing parameters.** We aim at finding the parameters that solve the partial factorization problem given $n$ LSBs of a prime $p$. We are interested in obtaining this parametrization for different values of $n$, starting from Coppersmith bound for RSA-2048: 512 to 552 bits. The optimization process is as follows: (1) for each $n$, $X$, $m$ and $t$ we generate 100 RSA-2048 keys using OpenSSL and try to recover the remaining bits of both primes; (2) filter out those sets ($X$, $m$, $t$) that do not achieve 100% success rate; (3) choose the set that minimizes $m + t$ for each $n$.

It is worth noting that each lattice test implies recovering the same key with $p$ and with $q$. This is due to the fact that in our attack scenario (see Section 3.3), the adversary is unaware if the known LSBs correspond to the larger or smaller prime, hence does not know if the bound is respected.

For all values of $n$, the bound $X = \frac{\sqrt{N}}{2^n}$ provides highly probable solutions and sometimes the lattice dimension shrinks by one. At first glance, this lattice reduction might seem insignificant. But, for example, with $n = 522$ it implies a runtime reduction of roughly 40 s. Indeed this is quick for a single lattice run, but when the number of candidates to test is high (i.e. after error correction) every second counts.

[5] D. Wong, function `coppersmith_howgrave_univariate` [link]

Table 4 summarizes the results of this characterization process for $X = \frac{\sqrt{N}}{2^n}$. One important aspect is that, for either value of $X$, we could not achieve the Coppersmith bound ($n = 512$). However, as pointed out in Section 5.1 our simulations suggest that the probability of recovering/correcting 512 bits is roughly the same as 522 bits, i.e. $n = 522$ is adequate in our setting.

We executed this parametrization on Sage 8.1 running on Ubuntu 16.04 on an Intel i7-3770 3.4 GHz. The running times in Table 4 correspond to the average of 100 lattice runs, dominated by the execution of LLL (Sage 8.1 default).

**Table 4:** Lattice attack characterization.

| $n$ | $m$ | $t$ | time (s) |
|---|---|---|---|
| 522 | 26 | 27 | 133.0 |
| 532 | 13 | 14 | 3.0 |
| 542 | 9 | 10 | 0.7 |
| 552 | 6 | 7 | 0.2 |