Homomorphic Evaluation of the AES Circuit

Craig Gentry
IBM Research

Shai Halevi
IBM Research

Nigel P. Smart
University of Bristol

February 26, 2012

Abstract

We describe a working implementation of leveled homomorphic encryption (without bootstrapping) that can evaluate the AES-128 circuit in three different ways. One variant takes under two days to evaluate an entire AES encryption operation, using NTL (over GMP) as our underlying software platform, and running on a large-memory machine. Using SIMD techniques, we can process over 54 blocks in each evaluation, yielding an amortized rate of just under one hour per block. Another implementation takes around five days to evaluate the AES operation, but can process 720 blocks in each evaluation, yielding an amortized rate of just over 10 minutes per block. We also detail a third implementation, which looks theoretically more attractive, but which in practice turns out to be less competitive.

For our implementations we develop both AES-specific optimizations as well as several "generic" tools for FHE evaluation. These last tools include (among others) a different variant of the Brakerski-Vaikuntanathan key-switching technique that does not require reducing the norm of the ciphertext vector, and a method of implementing the Brakerski-Gentry-Vaikuntanathan modulus-switching transformation on ciphertexts in CRT representation.

Keywords. AES, Fully Homomorphic Encryption, Implementation

The first and second authors are sponsored by DARPA under agreement number FA8750-11-C-0096. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

The third author is sponsored by DARPA and AFRL under agreement number FA8750-11-2-0079. The same disclaimers as above apply. He is also supported by the European Commission through the ICT Programme under Contract ICT-2007-216676 ECRYPT II and via an ERC Advanced Grant ERC-2010-AdG-267188-CRIPTO, by EPSRC via grant COED–EP/I03126X, and by a Royal Society Wolfson Merit Award. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the European Commission or EPSRC.

Contents

1 Introduction **1**

2 Background **2**
2.1 Notations and Mathematical Background . . . 2
2.2 BGV-type Cryptosystems . . . 3
2.3 Computing on Packed Ciphertexts . . . 5

3 General-Purpose Optimizations **6**
3.1 A New Variant of Key Switching . . . 6
3.2 Modulus Switching in Evaluation Representation . . . 7
3.3 Dynamic Noise Management . . . 8
3.4 Randomized Multiplication by Constants . . . 8

4 Homomorphic Evaluation of AES **9**
4.1 Homomorphic Evaluation of the Basic Operations . . . 9
4.2 Implementing The Permutations . . . 11
4.3 Performance Details . . . 12

References **12**

A More Details **14**
A.1 Plaintext Slots . . . 14
A.2 Canonical Embedding Norm . . . 14
A.3 Double CRT Representation . . . 15
A.4 Sampling From A_q . . . 16
A.5 Canonical embedding norm of random polynomials . . . 16

B The Basic Scheme **17**
B.1 Our Moduli Chain . . . 17
B.2 Modulus Switching . . . 17
B.3 Key Switching . . . 18
B.4 Key-Generation, Encryption, and Decryption . . . 20
B.5 Homomorphic Operations . . . 21

C Security Analysis and Parameter Settings **22**
C.1 Lower-Bounding the Dimension . . . 22
C.1.1 LWE with Sparse Key . . . 23
C.2 The Modulus Size . . . 24
C.3 Putting It Together . . . 26

D Further AES Implementation Methods **27**

E $\mathsf{Scale}(c, q_t, q_{t-1})$ **in** dble-CRT **Representation** **28**

1 Introduction

In his breakthrough result [11], Gentry demonstrated that fully-homomorphic encryption was theoretically possible, assuming the hardness of some problems in integer lattices. Since then, many different improvements have been made, for example authors have proposed new variants, improved efficiency, suggested other hardness assumptions, etc. Some of these works were accompanied by implementation [20, 12, 7, 21, 16], but all the implementations so far were either "proofs of concept" that can compute only one basic operation at a time (at great cost), or special-purpose implementations limited to evaluating very simple functions. In this work we report on the first implementation powerful enough to support an "interesting real world circuit". Specifically, we implemented a variant of the leveled FHE-without-bootstrapping scheme of Brakerski, Gentry, and Vaikuntanathan [4] (BGV), with support for deep enough circuits so that we can evaluate an entire AES-128 encryption operation.

Why AES? We chose to shoot for an evaluation of AES since it seems like a natural benchmark: AES is widely deployed and used extensively in security-aware applications (so it is "practically relevant" to implement it), and the AES circuit is nontrivial on one hand, but on the other hand not astronomical. Moreover the AES circuit has a regular (and quite "algebraic") structure , which is amenable to parallelism and optimizations. Indeed, for these same reasons AES is often used as a benchmark for implementations of protocols for secure multi-party computation (MPC), for example [19, 8, 14, 15]. Using the same yardstick to measure FHE and MPC protocols is quite natural, since these techniques target similar application domains and in some cases both techniques can be used to solve the same problem.

Beyond being a natural benchmark, homomorphic evaluation of AES decryption also has interesting applications: When data is encrypted under AES and we want to compute on that data, then homomorphic AES decryption would transform this AES-encrypted data into an FHE-encrypted data, and then we could perform whatever computation we wanted. (Such applications were alluded to in [16, 21]).

Our Contributions. Our implementation is based on a variant of the ring-LWE scheme of BGV [4, 6, 5], using the techniques of Smart and Vercauteren (SV) [21] and Gentry, Halevi and Smart (GHS) [13], and we introduce many new optimizations. Some of our optimizations are specific to AES, these are described in Section 4. Most of our optimization, however, are more general-purpose and can be used for homomorphic evaluation of other circuits, these are described in Section 3.

Many of our general-purpose optimizations are aimed at reducing the number of FFTs and CRTs that we need to perform, by reducing the number of times that we need to convert polynomials between coefficient and evaluation representations. Since the cryptosystem is defined over a polynomial ring, many of the operations involve various manipulation of integer polynomials, such as modular multiplications and additions and Frobenius maps. Most of these operations can be performed more efficiently in evaluation representation, when a polynomial is represented by the vector of values that it assumes in all the roots of the ring polynomial (for example polynomial multiplication is just point-wise multiplication of the evaluation values). On the other hand some operations in BGV-type cryptosystems (such as key switching and modulus switching) seem to require coefficient representation, where a polynomial is represented by listing all its coefficients.[1] Hence a "naive implementation" of FHE would need to convert the polynomials back and forth between the two representations, and these conversions turn out to be the most time-consuming part of the execution. In our implementation we keep ciphertexts in evaluation representation at all times, converting to coefficient representation only when needed for some operation, and then converting back.

[1] The need for coefficient representation ultimately stems from the fact that the noise in the ciphertexts is small in coefficient representation but not in evaluation representation.

We describe variants of key switching and modulus switching that can be implemented while keeping almost all the polynomials in evaluation representation. Our key-switching variant has another advantage, in that it significantly reduces the size of the key-switching matrices in the public key. This is particularly important since the main limiting factor for evaluating deep circuits turns out to be the ability to keep the key-switching matrices in memory. Other optimizations that we present are meant to reduce the number of modulus switching and key switching operations that we need to do. This is done by tweaking some operations (such as multiplication by constant) to get a slower noise increase, by "batching" some operations before applying key switching, and by attaching to each ciphertext an estimate of the "noisiness" of this ciphertext, in order to support better noise bookkeeping.

Our Implementation. Our implementation was based on the NTL C++ library running over GMP, we utilized a machine which consisted of a processing unit of Intel Xeon CPUs running at 2.0 GHz with 18MB cache, and most importantly with 256GB of RAM.[2] Memory was our main limiting factor in the implementation. With this machine it took us just under two days to compute a single block AES encryption using an implementation choice which minimizes the amount of memory required; this is roughly two orders of magnitude faster than what could be done with the Gentry-Halevi implementation [12]. The computation was performed on ciphertexts that could hold 864 plaintext slots each; where each slot holds an element of $\mathbb{F}_{2^8}$. This means that we can compute $\lfloor 864/16 \rfloor = 54$ AES operations in parallel, which gives an amortize time per block of roughly just under one hour. A second (byte-sliced) implementation, requiring more memory, completed an AES operation in around five days; where ciphertexts could now hold 720 different $\mathbb{F}_{2^8}$ slots. This results in an anortized time per block of roughly ten minutes.

We note that there are a multitude of optimizations that one can perform on our basic implementation. Most importantly, we believe that by using the "bootstrapping as optimization" technique from BGV [4] we can speedup the AES performance by an additional order of magnitude. Also, there are great gains to be had by making better use of parallelism: Unfortunately, the NTL library (which serves as our underlying software platform) is not thread safe, which severely limits our ability to utilize the multi-core functionality of modern processors (our test machine has 24 cores). We expect that by utilizing many threads we can speed up some of our (higher memory) AES variants by as much as a 16x factor; just by letting each thread compute a different S-box lookup.

Organization. In Section 2 we review the main features of BGV-type cryptosystems [5, 4], and briefly survey the techniques for homomorphic computation on packed ciphertexts from SV and GHS [21, 13]. Then in Section 3 we describe our "general-purpose" optimizations on a high level, with additional details provided in Appendices A and B. A brief overview of AES and a high-level description (and performance numbers) of one of our AES-specific implementations is provided in Section 4, with details of alternative implementations being provided in Appendix D.

2 Background

2.1 Notations and Mathematical Background

For an integer q we identify the ring $\mathbb{Z}/q\mathbb{Z}$ with the interval $(-q/2, q/2] \cap \mathbb{Z}$, and we use $[z]_q$ to denote the reduction of the integer z modulo q into that interval. Our implementation utilizes polynomial rings defined by cyclotomic polynomials, $\mathbb{A} = \mathbb{Z}[X]/\Phi_m(X)$. The ring $\mathbb{A}$ is the ring of integers of a the mth cyclotomic

[2]This machine was BlueCrystal Phase 2; and the authors would like to thank the University of Bristol's Advanced Computing Research Centre (`https://www.acrc.bris.ac.uk/`) for access to this facility

number field $\mathbb{Q}(\zeta_m)$. We let $\mathbb{A}_q \stackrel{\text{def}}{=} \mathbb{A}/q\mathbb{A} = \mathbb{Z}[X]/(\Phi_m(X), q)$ for the (possibly composite) integer q, and we identify $\mathbb{A}_q$ with the set of integer polynomials of degree upto $\phi(m) - 1$ reduced modulo q.

Coefficient vs. Evaluation Representation. Let m, q be two integers such that $\mathbb{Z}/q\mathbb{Z}$ contains a primitive m-th root of unity, and denote one such primitive m-th root of unity by $\zeta \in \mathbb{Z}/q\mathbb{Z}$. Recall that the m'th cyclotomic polynomial splits into linear terms modulo q, $\Phi_m(X) = \prod_{i \in (\mathbb{Z}/m\mathbb{Z})^*} (X - \zeta^i) \pmod q$.

For an element $a \in \mathbb{A}_q$, we consider two ways of representing it: Viewing a as a degree-$(\phi(m)-1)$ polynomial, $a(X) = \sum_{i<\phi(m)} a_i X^i$, we can just list all the coefficients in order $\mathbf{a} = \langle a_0, a_1, \ldots, a_{\phi(m)-1} \rangle \in (\mathbb{Z}/q\mathbb{Z})^{\phi(m)}$. We call $\mathbf{a}$ the *coefficient representation* of a. For the other representation we consider the values that the polynomial $a(X)$ assumes on all primitive m-th roots of unity modulo q, $b_i = a(\zeta^i) \bmod q$ for $i \in (\mathbb{Z}/m\mathbb{Z})^*$. The b_i's in order also yield a vector $\mathbf{b} \in (\mathbb{Z}/q\mathbb{Z})^{\phi(m)}$, which we call the *evaluation representation* of a. Clearly these two representations are related via $\mathbf{b} = V_m \cdot \mathbf{a}$, where V_m is the Vandermonde matrix over the primitive m-th roots of unity modulo q. We remark that for all i we have the equality $(a \bmod (X - \zeta^i)) = a(\zeta^i) = b_i$, hence the evaluation representation of a is just a polynomial Chinese-Remaindering representation.

In both evaluation and coefficient representations, an element $a \in \mathbb{A}_q$ is represented by a $\phi(m)$-vector of integers in $\mathbb{Z}/q\mathbb{Z}$. If q is a composite then each of these integers can itself be represented either using the standard binary encoding of integers or using Chinese-Remaindering relative to the factors of q. We usually use the standard binary encoding for the coefficient representation and Chinese-Remaindering for the evaluation representation. (Hence the latter representation is really a *double CRT* representation, relative to both the polynomial factors of $\Phi_m(X)$ and the integer factors of q.)

2.2 BGV-type Cryptosystems

Our implementation uses a variant of the BGV cryptosystem due to Gentry, Halevi and Smart, specifically the one described in [13, Appendix D] (in the full version). In this cryptosystem both ciphertexts and secret keys are vectors over the polynomial ring $\mathbb{A}$, and the native plaintext space is the space of binary polynomials $\mathbb{A}_2$. (More generally it could be $\mathbb{A}_p$ for some fixed $p \geq 2$, but in our case we will always use $\mathbb{A}_2$.)

At any point during the homomorphic evaluation there is some "current integer modulus q" and "current secret key $\mathbf{s}$", that change from time to time. A ciphertext $\mathbf{c}$ is decrypted using the current secret key $\mathbf{s}$ by taking inner product over $\mathbb{A}_q$ (with q the current modulus) and then reducing the result modulo 2 *in coefficient representation*. Namely, the decryption formula is

$$a \leftarrow [\ \underbrace{[\langle \mathbf{c}, \mathbf{s} \rangle \bmod \Phi_m(X)]_q}_{\text{noise}}\ ]_2 \,. \tag{1}$$

The polynomial $[\langle \mathbf{c}, \mathbf{s} \rangle \bmod \Phi_m(X)]_q$ is called the "noise" in the ciphertext $\mathbf{c}$. Informally, $\mathbf{c}$ is a *valid ciphertext* with respect to secret key $\mathbf{s}$ and modulus q if this noise has "sufficiently small norm" relative to q. The meaning of "sufficiently small norm" is whatever is needed to ensure that the noise does not wrap around q when performing homomorphic operations, in our implementation we keep the norm of the noise always below some pre-set bound (which is determined in Appendix C.2).

The specific norm that we use to evaluate the magnitude of the noise is the "canonical embedding norm reduced mod q", as described in [13, Appendix D] (in the full version). This is useful to get smaller parameters, but for the purpose of presentation the reader can think of the norm as the Euclidean norm of the noise in coefficient representation. More details are given in the Appendices. We refer to the norm of the noise as *the noise magnitude*.

The central feature of BGV-type cryptosystems is that the current secret key and modulus evolve as we apply operations to ciphertexts. We apply five different operations to ciphertexts during homomorphic evaluation. Three of them — addition, multiplication, and automorphism — are "semantic operations" that we use to evolve the plaintext data which is encrypted under those ciphertexts. The other two operations — key-switching and modulus-switching — are used for "maintenance": These operations do not change the plaintext at all, they only change the current key or modulus (respectively), and they are mainly used to control the complexity of the evaluation. Below we briefly describe each of these five operations on a high level. For the sake of self-containment, we also describe key generation and encryption in Appendix B. More detailed description can be found in [13, Appendix D].

Addition. Homomorphic addition of two ciphertext vectors with respect to the same secret key and modulus q is done just by adding the vectors over $\mathbb{A}_q$. If the two arguments were encrypting the plaintext polynomials $a_1, a_2 \in \mathbb{A}_2$ then the sum will be an encryption of $a_1 + a_2 \in \mathbb{A}_2$. This operation has no effect on the current modulus or key, and the norm of the noise is at most the sum of norms from the noise in the two arguments.

Multiplication. Homomorphic multiplication is done via tensor product over $\mathbb{A}_q$. In principle, if the two arguments have dimension n over $\mathbb{A}_q$ then the product ciphertext has dimension n^2, each entry in the output computed as the product of one entry from the first argument and one entry from the second.[3]

This operation does not change the current modulus, but it changes the current key: If the two input ciphertexts are valid with respect to the dimension-n secret key vector $\mathbf{s}$, encrypting the plaintext polynomials $a_1, a_2 \in \mathbb{A}_2$, then the output is valid with respect to the dimension-n^2 secret key $\mathbf{s}'$ which is the tensor product of $\mathbf{s}$ with itself, and it encrypt the polynomial $a_1 \cdot a_2 \in \mathbb{A}_2$. The norm of the noise in the product ciphertext can be bounded in terms of the product of norms of the noise in the two arguments. For our choice of norm function, the norm of the product is no larger than the product of the norms of the two arguments.

Key Switching. The public key of BGV-type cryptosystems includes additional components to enable converting a valid ciphertext with respect to one key into a valid ciphertext encrypting the same plaintext with respect to another key. For example, this is used to convert the product ciphertext which is valid with respect to a high-dimension key back to a ciphertext with respect to the original low-dimension key.

To allow conversion from dimension-n' key $\mathbf{s}'$ to dimension-n key $\mathbf{s}$ (both with respect to the same modulus q), we include in the public key a matrix $W = W[\mathbf{s}' \to \mathbf{s}]$ over $\mathbb{A}_q$, where the i'th column of W is roughly an encryption of the i'th entry of $\mathbf{s}'$ with respect to $\mathbf{s}$ (and the current modulus). Then given a valid ciphertext $\mathbf{c}'$ with respect to $\mathbf{s}'$, we roughly compute $\mathbf{c} = W \cdot \mathbf{c}'$ to get a valid ciphertext with respect to $\mathbf{s}$.

In some more detail, the BGV key switching transformation first ensures that the norm of the ciphertext $\mathbf{c}'$ itself is sufficiently low with respect to q. In [4] this was done by working with the binary encoding of $\mathbf{c}'$, and one of our main optimization in this work is a different method for achieving the same goal (cf. Section 3.1). Then, if the i'th entry in $\mathbf{s}'$ is $\mathfrak{s}'_i \in \mathbb{A}$ (with norm smaller than q), then the i'th column of $W[\mathbf{s}' \to \mathbf{s}]$ is an n-vector $\mathbf{w}_i$ such that $[\langle \mathbf{w}_i, \mathbf{s} \rangle \bmod \Phi_m(X)]_q = 2e_i + \mathfrak{s}'_i$ for a low-norm polynomial $e_i \in \mathbb{A}$. Denoting $\mathbf{e} = (e_1, \ldots, e_{n'})$, this means that we have $\mathbf{s}W = \mathbf{s}' + 2\mathbf{e}$ over $\mathbb{A}_q$. For any ciphertext vector $\mathbf{c}'$, setting $\mathbf{c} = W \cdot \mathbf{c}' \in \mathbb{A}_q$ we get the equation

$$[\langle \mathbf{c}, \mathbf{s} \rangle \bmod \Phi_m(X)]_q = [\mathbf{s}W\mathbf{c}' \bmod \Phi_m(X)]_q = [\langle \mathbf{c}', \mathbf{s}' \rangle + 2\langle \mathbf{c}', \mathbf{e} \rangle \bmod \Phi_m(X)]_q$$

Since $\mathbf{c}'$, $\mathbf{e}$, and $[\langle \mathbf{c}', \mathbf{s}' \rangle \bmod \Phi_m(X)]_q$ all have low norm relative to q, then the addition on the right-hand side does not cause a wrap around q, hence we get $[[\langle \mathbf{c}, \mathbf{s} \rangle \bmod \Phi_m(X)]_q]_2 = [[\langle \mathbf{c}', \mathbf{s}' \rangle \bmod \Phi_m(X)]_q]_2$, as

[3] It was shown in [6] that over polynomial rings this operation can be implemented while increasing the dimension only to $2n-1$ rather than to n^2.

needed. The key-switching operation changes the current secret key from $\mathbf{s}'$ to $\mathbf{s}$, and does not change the current modulus. The norm of the noise is increased by at most an additive factor of $2\| \langle \mathbf{c}', \mathbf{e} \rangle \|$.

Modulus Switching. The modulus switching operation is intended to reduce the norm of the noise, to compensate for the noise increase that results from all the other operations. To convert a ciphertext $\mathbf{c}$ with respect to secret key $\mathbf{s}$ and modulus q into a ciphertext $\mathbf{c}'$ encrypting the same thing with respect to the same secret key but modulus q', we roughly just scale $\mathbf{c}$ by a factor q'/q (thus getting a fractional ciphertext), then round appropriately to get back an integer ciphertext. Specifically $\mathbf{c}'$ is a ciphertext vector satisfying (a) $\mathbf{c}' = \mathbf{c} \pmod 2$, and (b) the "rounding error term" $\tau \stackrel{\text{def}}{=} \mathbf{c}' - (q'/q)\mathbf{c}$ has low norm. Converting $\mathbf{c}$ to $\mathbf{c}'$ is easy in coefficient representation, and one of our optimizations is a method for doing the same in evaluation representation (cf. Section 3.2) This operation leaves the current key $\mathbf{s}$ unchanged, changes the current modulus from q to q', and the norm of the noise is changed as $\|n'\| \leq (q'/q)\|n\| + \|\tau \cdot s\|$. Note that if the key $\mathbf{s}$ has low norm and q' is sufficiently smaller than q, then the noise magnitude decreases by this operation.

A BGV-type cryptosystem has a chain of moduli, $q_0 < q_1 \cdots < q_{L-1}$, where fresh ciphertexts are with respect to the largest modulus q_{L-1}. During homomorphic evaluation every time the (estimated) noise grows too large we apply modulus switching from q_i to q_{i-1} in order to decrease it back. Eventually we get ciphertexts with respect to the smallest modulus q_0, and we cannot compute on them anymore (except by using bootstrapping).

Automorphisms. In addition to adding and multiplying polynomials, another useful operation is converting the polynomial $a(X) \in \mathbb{A}$ to $a^{(i)}(X) \stackrel{\text{def}}{=} a(X^i) \bmod \Phi_m(X)$. Denoting by κ_i the transformation $\kappa_i : a \mapsto a^{(i)}$, it is a standard fact that the set of transformations $\{\kappa_i : i \in (\mathbb{Z}/m\mathbb{Z})^*\}$ forms a group under composition (which is the Galois group $\mathcal{G}\mathsf{al}(\mathbb{Q}(\zeta_m)/\mathbb{Q})$), and this group is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^*$. In [4, 13] it was shown that applying the transformations κ_i to the plaintext polynomials is very useful, some more examples of its use can be found in our Section 4.

Denoting by $\mathbf{c}^{(i)}$, $\mathbf{s}^{(i)}$ the vector obtained by applying κ_i to each entry in $\mathbf{c}, \mathbf{s}$, respectively, it was shown in [4, 13] that if $\mathbf{s}$ is a valid ciphertext encrypting a with respect to key $\mathbf{s}$ and modulus q, then $\mathbf{c}^{(i)}$ is a valid ciphertext encrypting $a^{(i)}$ with respect to key $\mathbf{s}^{(i)}$ and the same modulus q. Moreover the norm of noise remains the same under this operation. We remark that we can apply key-switching to $\mathbf{c}^{(i)}$ in order to get an encryption of $a^{(i)}$ with respect to the original key $\mathbf{s}$.

2.3 Computing on Packed Ciphertexts

Smart and Vercauteren observed [20, 21] that the plaintext space $\mathbb{A}_2$ can be viewed as a vector of "plaintext slots", by an application the polynomial Chinese Remainder Theorem. Specifically, if the ring polynomial $\Phi_m(X)$ factors modulo 2 into a product of irreducible factors $\Phi_m(X) = \prod_{j=0}^{\ell-1} F_j(X) \pmod 2$, then a plaintext polynomial $a(X) \in \mathbb{A}_2$ can be viewed as encoding ℓ different small polynomials, $a_j = a \bmod F_j$. Just like for integer Chinese Remaindering, addition and multiplication in $\mathbb{A}_2$ correspond to element-wise addition and multiplication of the vectors of slots.

The effect of the automorphisms is a little more involved. When i is a power of two then the transformations $\kappa_i : a \mapsto a^{(i)}$ is just applied to each slot separately. When i is not a power of two the transformation κ_i has the effect of roughly shifting the values between the different slots. For example, for some parameters we could get a cyclic shift of the vector of slots: If a encodes the vector $(a_0, a_1, \ldots, a_{\ell-1})$, then $\kappa_i(a)$ (for some i) could encode the vector $(a_{\ell-1}, a_0, \ldots, a_{\ell-2})$. This was used in [13] to devise efficient procedures for applying arbitrary permutations to the plaintext slots.

We note that the values in the plaintext slots are not just bits, rather they are polynomials modulo the irreducible F_j's, so they can be used to represents elements in extension fields GF(2^d). In particular, in some of our AES implementations we used the plaintext slots to hold elements of GF(2^8), and encrypt one byte of the AES state in each slot. Then we can use an adaption of the techniques from [13] to permute the slots when performing the AES row-shift and column-mix.

3 General-Purpose Optimizations

Below we summarize our optimizations that are not tied directly to the AES circuit and can be used also in homomorphic evaluation of other circuits. Underlying many of these optimizations is our choice of keeping ciphertext and key-switching matrices in evaluation (double-CRT) representation. Our chain of moduli is defined via a set of primes of roughly the same size, $p_0, \ldots, p_{L-1}$, all chosen such that $\mathbb{Z}/p_i\mathbb{Z}$ has a m'th roots of unity. (In other words, $m|p_i - 1$ for all i.) For $i = 0, \ldots, L-1$ we then define our i'th modulus as $q_i = \prod_{j=0}^{i} p_i$. The primes p_0 and p_{L-1} are special (p_0 is chosen to ensure decryption works, and p_{L-1} is chosen to control noise immediately after encryption), however all other primes p_i are of size $2^{17} \leq p_i \leq 2^{20}$ if $L < 100$, see Appendix C.

In the t-th level of the scheme we have ciphertexts consisting of elements in $\mathbb{A}_{q_t}$ (i.e., polynomials modulo $(\Phi_m(X), q_t)$). We represent an element $c \in \mathbb{A}_{q_t}$ by a $\phi(m) \times (t+1)$ "matrix" of its evaluations at the primitive m-th roots of unity modulo the primes $p_0, \ldots, p_t$. Computing this representation from the coefficient representation of c involves reducing c modulo the p_i's and then $t+1$ invocations of the FFT algorithm, modulo each of the p_i (picking only the FFT coefficients corresponding to $(\mathbb{Z}/m\mathbb{Z})^*$). To convert back to coefficient representation we invoke the inverse FFT algorithm $t+1$ times, each time padding the $\phi(m)$-vector of evaluation point with $m - \phi(m)$ zeros (for the evaluations at the non-primitive roots of unity). This yields the coefficients of $t+1$ polynomials modulo $(X^m - 1, p_i)$ for $i = 0, \ldots, t$, we then reduce each of these polynomials modulo $(\Phi_m(X), p_i)$ and apply Chinese Remainder interpolation. We stress that we try to perform these transformations as rarely as we can.

3.1 A New Variant of Key Switching

As described in Section 2, the key-switching transformation introduces an additive factor of $2\langle \mathbf{c}', \mathbf{e} \rangle$ in the noise, where $\mathbf{c}'$ is the input ciphertext and $\mathbf{e}$ is the noise component in the key-switching matrix. To keep the noise magnitude below the modulus q, it seems that we need to ensure that the ciphertext $\mathbf{c}'$ itself has low norm. In BGV [4] this was done by representing $\mathbf{c}'$ as a fixed linear combination of small vectors, i.e. $\mathbf{c}' = \sum_i 2^i \mathbf{c}'_i$ with $\mathbf{c}'_i$ the vector of i'th bits in $\mathbf{c}'$. Considering the high-dimension ciphertext $\mathbf{c}^* = (\mathbf{c}'_0|\mathbf{c}'_1|\mathbf{c}'_2|\cdots)$ and secret key $\mathbf{s}^* = (\mathbf{s}'|2\mathbf{s}'|4\mathbf{s}'|\cdots)$, we note that we have $\langle \mathbf{c}^*, \mathbf{s}^* \rangle = \langle \mathbf{c}', \mathbf{s}' \rangle$, and $\mathbf{c}^*$ has low norm (since it consists of 0-1 polynomials). BGV therefore included in the public key the matrix $W = W[\mathbf{s}^* \to \mathbf{s}]$ (rather than $W[\mathbf{s}' \to \mathbf{s}]$), and had the key-switching transformation computes $\mathbf{c}^*$ from $\mathbf{c}'$ and sets $\mathbf{c} = W \cdot \mathbf{c}^*$.

When implementing key-switching, there are two drawbacks to the above approach. First, this increases the dimension (and hence the size) of the key switching matrix. This drawback is fatal when evaluating deep circuits, since having enough memory to keep the key-switching matrices turns out to be the limiting factor in our ability to evaluate these deep circuits. Another drawback is it seems that this key-switching procedure requires that we first convert $\mathbf{c}'$ to coefficient representation in order to compute the $\mathbf{c}'_i$'s, then convert each of the $\mathbf{c}'_i$'s back to evaluation representation before multiplying by the key-switching matrix. In level t of the circuit, this seem to require $\Omega(t \log q_t)$ FFTs.

In this work we propose a different variant: Rather than manipulating $\mathbf{c}'$ to decrease its norm, we instead temporarily increase the modulus q. To that end we recall that for a valid ciphertext $\mathbf{c}'$, encrypting plaintext a with respect to $\mathbf{s}'$ and q, we have the equality $\langle \mathbf{c}', \mathbf{s}' \rangle = 2e' + a$ over A_q, for a low-norm polynomial e'. This equality, we note, implies that for every odd integer p we have the equality $\langle \mathbf{c}', p\mathbf{s}' \rangle = 2e'' + a$, *holding over* A_{pq}, for the "low-norm" polynomial e'' (namely $e'' = p \cdot e' + \frac{p-1}{2} a$). Clearly, when considered relative to secret key $p\mathbf{s}$ and modulus pq, the noise in $\mathbf{c}'$ is p times larger than it was relative to $\mathbf{s}$ and q. However, since the modulus is also p times larger, we maintain that the noise has norm sufficiently smaller than the modulus. In other words, $\mathbf{c}'$ is still a valid ciphertext that encrypts the same plaintext a with respect to secret key $p\mathbf{s}$ and modulus pq. By taking p large enough, we can ensure that the norm of $\mathbf{c}'$ (which is independent of p) is sufficiently small relative to the modulus pq.

We therefore include in the public key a matrix $W = W[p\mathbf{s}' \to \mathbf{s}]$ modulo pq for a large enough odd integer p. (Specifically we need $p \approx q\sqrt{m}$.) Given a ciphertext $\mathbf{c}'$, valid with respect to $\mathbf{s}$ and q, we apply the key-switching transformation simply by setting $\mathbf{c} = W \cdot \mathbf{c}'$ over $\mathbb{A}_{pq}$. The additive noise term $\langle \mathbf{c}', \mathbf{e} \rangle$ that we get is now small enough relative to our large modulus pq, thus the resulting ciphertext $\mathbf{c}$ is valid with respect to $\mathbf{s}$ and pq. We can now switch the modulus back to q (using our modulus switching routine), hence getting a valid ciphertext with respect to $\mathbf{s}$ and q.

We note that even though we no longer break $\mathbf{c}'$ into its binary encoding, it seems that we still need to recover it in coefficient representation in order to compute the evaluations of $\mathbf{c}' \bmod p$. However, since we do not increase the dimension of the ciphertext vector, this procedure requires only $O(t)$ FFTs in level t (vs. $O(t \log q_t) = O(t^2)$ for the original BGV variant). Also, the size of the key-switching matrix is reduced by roughly the same factor of $\log q_t$.

Our new variant comes with a price tag, however: We use key-switching matrices relative to a larger modulus, but still need the noise term in this matrix to be small. This means that the LWE problem underlying this key-switching matrix has larger ratio of modulus/noise, implying that we need a larger dimension to get the same level of security than with the original BGV variant. In fact, since our modulus is more than squared (from q to pq with $p > q$), the dimension is increased by more than a factor of two. This translates to more than doubling of the key-switching matrix, partly negating the size and running time advantage that we get from this variant.

We comment that a hybrid of the two approaches could also be used: we can decrease the norm of $\mathbf{c}'$ only somewhat by breaking it into digits (as opposed to binary bits as in [4]), and then increase the modulus somewhat until it is large enough relative to the smaller norm of $\mathbf{c}'$. We speculate that the optimal setting in terms of runtime is found around $p \approx \sqrt{q}$, but so far did not try to explore this tradeoff.

3.2 Modulus Switching in Evaluation Representation

Given an element $c \in \mathbb{A}_{q_t}$ in evaluation (double-CRT) representation relative to $q_t = \prod_{j=0}^{t} p_j$, we want to modulus-switch to q_{t-1} – i.e., scale down by a factor of p_t; we call this operation $\mathsf{Scale}(c, q_t, q_{t-1})$ The output should be $c' \in \mathbb{A}$, represented via the same double-CRT format (with respect to $p_0, \ldots, p_{t-1}$), such that (a) $c' \equiv c \pmod 2$, and (b) the "rounding error term" $\tau = c' - (c/p_t)$ has a very low norm. As p_t is odd, we can equivalently require that the element $c^\dagger \stackrel{\text{def}}{=} p_t \cdot c'$ satisfy

(i) $c^\dagger$ is divisible by p_t,

(ii) $c^\dagger \equiv c \pmod 2$, and

(iii) $c^\dagger - c$ (which is equal to $p_t \cdot \tau$) has low norm.

Rather than computing c' directly, we will first compute $c^\dagger$ and then set $c' \leftarrow c^\dagger/p_t$. Observe that once we compute $c^\dagger$ in double-CRT format, it is easy to output also c' in double-CRT format: given the evaluations for $c^\dagger$ modulo p_j ($j < t$), simply multiply them by $p_t^{-1} \bmod p_j$. The algorithm to output $c^\dagger$ in double-CRT format is as follows:

1. Set $\bar{c}$ to be the coefficient representation of $c \bmod p_t$. (Computing this requires a single "small FFT" modulo the prime p_t.)
2. Add or subtract p_t from every odd coefficient of $\bar{c}$, thus obtaining a polynomial δ with coefficients in $(-p_t, p_t]$ such that $\delta \equiv \bar{c} \equiv c \pmod{p_t}$ and $\delta \equiv 0 \pmod{2}$.
3. Set $c^\dagger = c - \delta$, and output it in double-CRT representation.

   Since we already have c in double-CRT representation, we only need the double-CRT representation of δ, which requires t more "small FFTs" modulo the p_j's.

As all the coefficients of $c^\dagger$ are within p_t of those of c, the "rounding error term" $\tau = (c^\dagger - c)/p_t$ has coefficients of magnitude at most one, hence it has low norm.

The procedure above uses $t+1$ small FFTs in total. This should be compared to the naive method of just converting everything to coefficient representation modulo the primes ($t+1$ FFTs), CRT-interpolating the coefficients, dividing the and rounding appropriately the large integers (of size $\approx q_t$), CRT-decomposing the coefficients, and then converting back to evaluation representation ($t+1$ more FFTs). The above approach makes explicit use of the fact that we are working in a plaintext space modulo 2; in Appendix E we present a technique which works when the plaintext space is defined modulo a larger modulus.

3.3 Dynamic Noise Management

As described in the literature, BGV-type cryptosystems tacitly assume that each homomorphic operation operation is followed a modulus switch to reduce the noise magnitude. In our implementation, however, we attach to each ciphertext an estimate of the noise magnitude in that ciphertext, and use these estimates to decide dynamically when a modulus switch must be performed.

Each modulus switch consumes a level, and hence a goal is to reduce, over a computation, the number of levels consumed. By paying particular attention to the parameters of the scheme, and by carefully analyzing how various operations affect the noise, we are able to control the noise much more carefully than in prior work. In particular, we note that modulus-switching is really only necessary just prior to multiplication (when the noise magnitude is about to get squared), in other times it is acceptable to keep the cipehrtexts at a higher level (with higher noise). In addition to get better parameters we measure the noise using a new measure based on norms in the canonical embedding.

3.4 Randomized Multiplication by Constants

Our implementation of the AES round function uses just a few multiplication operations (only seven per byte!), but it requires a relatively large number of multiplications of encrypted bytes by constants. Hence it becomes important to try and squeeze down the increase in noise when multiplying by a constant. To that end, we encode a constant polynomial in $\mathbb{A}_2$ as a polynomial with coefficients in $\{-1, 0, 1\}$ rather than in $\{0, 1\}$. Namely, we have a procedure $\mathsf{Randomize}(\alpha)$ that takes a polynomial $\alpha \in \mathbb{A}_2$ and replaces each non-zero coefficients with a coefficients chosen uniformly from $\{-1, 1\}$. By Chernoff bound, we expect that for α with h nonzero coefficients, the canonical embedding norm of $\mathsf{Randomize}(\alpha)$ to be bounded by

$O(\sqrt{h})$ with high probability (assuming that h is large enough for the bound to kick in). This yields a better bound on the noise increase than the trivial bound of h that we would get if we just multiply by α itself. (In Appendix A.5 we present a heuristic argument that we use to bound the noise, which yeilds the same asymptotic bounds but slightly better constants.)

4 Homomorphic Evaluation of AES

Next we describe our homomorphic implementation of AES-128. We implemented three distinct implementation possibilities; below we describe one of them, in which the entire AES state is packed in just one ciphertext. Two other implementations (of byte-slice and bit-slice AES) are described in Appendix D. The implementation described here uses the least amount of memory (which turns out to be the main constraint in our implementation), and also the fastest running time for a single evaluation. The other implementation choices allow more SIMD parallelism, on the other hand, so they can give better amortized running time when evaluating AES on many blocks in parallel.

A Brief Overview of AES. The AES-128 cipher consists of ten applications of the same keyed round function (with different round keys). The round function operates on a 4×4 matrix of bytes, which are sometimes considered as element of $\mathbb{F}_{2^8}$. The basic operations that are performed during the round function are AddKey, SubBytes, ShiftRows, MixColumns. The AddKey is simply an XOR operation of the current state with 16 bytes of key; the SubBytes operation consists of an inversion in the field $\mathbb{F}_{2^8}$ followed by a fixed $\mathbb{F}_2$-linear map on the bits of the element (relative to a fixed polynomial representation of $\mathbb{F}_{2^8}$); the ShiftRows rotates the entries in the row i of the 4×4 matrix by $i-1$ places to the left; finally the MixColumns operations pre-multiplies the state matrix by a fixed 4×4 matrix.

Our Packed Representation of the AES state. For our implementation we chose the native plaintext space of our homomorphic encryption so as to support operations on the finite field $\mathbb{F}_{2^8}$. To this end we choose our ring polynomial as $\Phi_m(X)$ that factors modulo 2 into degree-d irreducible polynomials such that $8|d$. (In other words, the smallest integer d such that $m|(2^d - 1)$ is divisible by 8.) This means that our plaintext slots can hold elements of $\mathbb{F}_{2^d}$, and in particular we can use them to hold elements of $\mathbb{F}_{2^8}$ which is a sub-field of $\mathbb{F}_{2^d}$. Since we have $\ell = \phi(m)/d$ plaintext slots in each ciphertext, we can represent upto $\lfloor \ell/16 \rfloor$ complete AES state matrices per ciphertext.

4.1 Homomorphic Evaluation of the Basic Operations

We now examine each AES operation in turn, and describe how it is implemented homomorphically. For each operation we denote the plaintext polynomial underlying a given input ciphertext $\mathfrak{c}$ by a, and the corresponding content of the ℓ plaintext slots are denoted as an ℓ-vector $(\alpha_i)_{i=1}^{\ell}$, with each $\alpha_i \in \mathbb{F}_{2^8}$.

AddKey. The AddKey is just a simple addition of ciphertexts.

SubBytes. We implement $\mathbb{F}_{2^8}$ inversion followed by the $\mathbb{F}_2$ affine transformation using the Frobenius automorphisms, $X \longrightarrow X^{2^j}$. Recall that for a power of two $k = 2^j$, the transformation $\kappa_k(a(X)) = (a(X^k) \bmod \Phi_m(X))$ is applied separately to each slot, hence we can use it to transform the vector $(\alpha_i)_{i=1}^{\ell}$ into $(\alpha_i^k)_{i=1}^{\ell}$. Inversion in $\mathbb{F}_{2^8}$ can be expressed as

$$\alpha^{-1} = \alpha^{254} = \left(\left(\alpha^{2^7} \cdot \alpha^{2^6}\right) \cdot \left(\alpha^{2^5} \cdot \alpha^{2^4}\right)\right) \cdot \left(\left(\alpha^{2^3} \cdot \alpha^{2^2}\right) \cdot \alpha^{2^1}\right),$$

which we implement by computing homomorphically the seven ciphertexts encrypting the polynomials $\kappa_k(a)$ for $k = 2, 4, \ldots, 128$, then using a depth-3 multiplication circuit to multiply these seven ciphertexts, thus getting an encryption of of a plaintext polynomial b that encodes the vector $(\alpha_i^{-1})_{i=1}^{\ell}$ in its slots.

To apply the AES $\mathbb{F}_2$ affine transformation, we use the fact that any $\mathbb{F}_2$ affine transformation can be computed as a $\mathbb{F}_{2^8}$ affine transformation over the conjugates. Thus there are constants $\gamma_0, \gamma_1, \ldots, \gamma_7, \delta \in \mathbb{F}_{2^8}$ such that the AES affine transformation $T_{\mathrm{AES}}(\cdot)$ can be expressed as $T_{\mathrm{AES}}(\beta) = \delta + \sum_{j=0}^{7} \gamma_j \cdot \beta^{2^j}$ over $\mathbb{F}_{2^8}$. We therefore again apply the Frobenius automorphisms to compute eight ciphertexts encrypting the polynomials $\kappa_k(b)$ for $k = 1, 2, 4, \ldots, 128$, and take the appropriate linear combination (with coefficients the γ_j's) to get an encryption of the vector $(T_{\mathrm{AES}}(\alpha_i^{-1}))_{i=1}^{\ell}$.

One subtle implementation detail to note here, is that although our plaintext slots all hold elements of the same field $\mathbb{F}_{2^8}$, they hold these elements with respect to different polynomial encodings. The AES affine transformation, on the other hand, is defined with respect to one particular fixed polynomial encoding. This means that we must implement in the i'th slot not the affine transformation $T_{\mathrm{AES}}(\cdot)$ itself but rather the projection of this transformation onto the appropriate polynomial encoding: When we take the affine transformation of the eight ciphertexts encrypting $b_j = \kappa_{2^j}(b)$, we therefore multiply the encryption of b_j not by a constant that has γ_j in all the slots, but rather by a constant that has in slot i the projection of γ_j to the polynomial encoding of slot i.

Implementing the procedure above, we note that applying the Frobenius automorphisms has almost no influence on the noise magnitude, and hence it does not consume any levels.[4] A multiplication operation consumes one full level of the circuit, and a multiplication-by-constant consumes roughly half a level in terms of added noise.[5] Below we provide a pseudo-code description of each stage, together with an approximation of the levels that are consumed by these operations. (These approximations are somewhat under-estimates, however.)

Input: ciphertext $\mathfrak{c}$ in level t		Level
// Compute $\mathfrak{c}_{254} = \mathfrak{c}^{-1}$		
1. $\mathfrak{c}_k \leftarrow \kappa_k(\mathfrak{c})$ for $k = 2, 4, 8, \ldots, 128$		t
2. $\mathfrak{c}_{254} \leftarrow ((\mathfrak{c}_{128} \cdot \mathfrak{c}_{64}) \cdot (\mathfrak{c}_{32} \cdot \mathfrak{c}_{16})) \cdot ((\mathfrak{c}_8 \cdot \mathfrak{c}_4) \cdot \mathfrak{c}_2)$	// Three multiplication levels	$t - 3.0$
// Affine transformation over $\mathbb{F}_2$		
3. $\mathfrak{c}'_k \leftarrow \kappa_k(\mathfrak{c}_{254})$ for $k = 1, 2, 4, 8, \ldots, 128$		$t - 3.0$
4. $\mathfrak{c}'' \leftarrow \delta + \sum_{j=0}^{7} \gamma_j \cdot \mathfrak{c}'_{2^j}$	// Linear combination over $\mathbb{F}_{2^8}$	$t - 3.5$

ShiftRows/MixColumns. As commonly done, we interleave the ShiftRows/MixColumns operations, viewing both as a single linear transformation over vectors from $(\mathbb{F}_{2^8})^{16}$. Given the ciphertext $\mathfrak{c}''$ after the SubBytes step, we compute from it four ciphertexts corresponding to permutations of the 16 bytes inside each of the $\ell/16$ different input blocks. Denoting the slots in one block by indexes $\{0, \ldots, 15\}$, we use the

[4] It does increase the noise magnitude somewhat, because we need to do key switching after these automorphisms. But this is only a small influence, and we will ignore it here.

[5] Recall though that we only switch levels on entry into a full multiplication operation; so the estimate of half a level is purely an approximation.

following four permutations (in disjoint cycle notation):

$$\pi_1 = (4,5,6,7)(8,10)(9,11)(15,14,13,12),$$
$$\pi_2 = (0,5,11,14,2,7,9,12)(1,6,8,15,3,4,10,13)$$
$$\pi_3 = (0,10,2,8)(1,11,3,9)(4,15)(5,12)(6,13)(7,14)$$
$$\pi_4 = (0,15,9,6,2,13,11,4)(1,12,10,7,3,14,8,5).$$

These four ciphertexts are combined via a linear operation (with coefficients $1, X$, and $(1+X)$) to obtain the final result of this round function. Below is a pseudo-code of this implementation and an approximation for the levels that it consumes (starting from $t-3.5$). We note that the permutations are implemented using automorphisms and multiplication by constant, thus we expect them to consume roughly 1/2 level. See Section 4.2 for more details.

Input: ciphertext $\mathfrak{c}''$ in level $t-3.5$	Level
5. $\mathfrak{c}^*_j \leftarrow \pi_j(\mathfrak{c}'')$ for $j=1,2,3,4$	$t-4.0$
6. Output $X\cdot\mathfrak{c}^*_1 + (X+1)\cdot\mathfrak{c}^*_2 + \mathfrak{c}^*_3 + \mathfrak{c}^*_4$	$t-4.5$

The Cost of One Round Function. The above description yields an estimate of 4.5 levels for implementing one round function. This is however, an underestimate. The actual number of levels depends on details such as how sparse the scalars are with respect to the embedding via Φ_m in a given parameter set, as well as the accumulation of noise with respect to additions, Frobenius operations etc. Running over many different parameter sets we find the average number of levels per round for this method varies between 5.0 and 6.0.

We mention that the byte-slice and bit-slice implementations, given in Appendix D, can consume less levels per round function, since they do not need to permute slots inside a single ciphertext. Specifically, for our byte-sliced implementation, we only need 4.5-5.0 levels per round on average. However, since we need to manipulate many more ciphertexts, the implementation takes much more time per evaluation and requires much more memory. On the other hand it offers wider parallelism, so yields better amortized time per block. Our bit-sliced implementation should theoretical consume the least number of levels (by purely counting multiplication gates), but the noise introduced by additions means the average number of levels consumed per round varies from 5.0 upto 10.0.

4.2 Implementing The Permutations

We now explain how to implement the permutations π_i that are needed for the ShiftRows/MixColumns step. Gentry et al. described in [13] how to implement an arbitrary permutation on the slots $\{0,1,\ldots,\ell-1\}$ in complexity only polylog(ℓ) using permutation networks of depth $O(\log\ell)$. We do not need arbitrary permutations, however, we only use very simple permutations that permute each block of 16 slots separately. This results in the permutation being highly regular in nature (i.e. they are composed of a combination of shift-left/shift-right operations of a given fixed amount). We can therefore use a simpler method for implementing these permutations, which has asymptotic complexity $O(\ell)$ for permuting ℓ slots (rather than polylog(ℓ)) but can be implemented in lower depth.

Recall that to implement permutation over the slots we use the maps κ_i ($X \mapsto X^i$) for indexes $1 < i < m$, which are not powers of two. Roughly, these maps implement a small set of "simple permutations" from which we can build arbitrary permutations as described in [13]. One useful property of these maps is that they are sharply transitive, namely for every j_1, j_2 there exists an index $i = i(j_1, j_2)$ such that κ_i moves the

content of slot j_1 into slot j_2. Below we abuse notations somewhat and denote that map by κ_{j_1,j_2} rather than $\kappa_{i(j_1,j_2)}$.

Given a ciphertext $\mathfrak{c}$, we can implement a permutation π over the slots $\{0,1,\ldots,15\}$ under this ciphertext by computing upto 16 maps $\mathfrak{c}_i \leftarrow \kappa_{i,\pi(i)}$ for $i = 0,1,\ldots,15$, then using a big MUX operations to select slot $\pi(i)$ from the ciphertext $\mathfrak{c}_i$. It was shown in [13] how to implement a two-input MUX using multiplication by a constant selection vector followed by addition (this operation is called Select in [13]). The same technique easily extend to larger MUXes. The end result is an implementation of the permutations π_i, consisting of one layer of Frobenius maps followed by one layer of multiplication by constants and then addition.

Low memory permutations. With the above method we need a large number of key-switching matrices in the public key, namely one for each of the maps κ_{j_1,j_2}'s. We can reduces this memory requirement, at the expense of taking longer to perform the permutations. We use the fact that the Galois group $\mathcal{G}\mathsf{al}$ that contains all the maps κ_i (which is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^*$) is generated by a relatively small number (roughly $\Theta(\log\log m)$) of generators Specifically, for our choice of parameters the group $(\mathbb{Z}/m\mathbb{Z})^*$ has two or three generators. It is therefore enough to store in the public key only the key-switching matrices corresponding to the generators of the group $\mathcal{G}\mathsf{al}$, then in order to apply a map κ_{j_1,j_2} we express it as a product of the generators and apply these generators to get the effect of κ_i. (For example, if $\kappa_{j_1,j_2} = g_1^2 \cdot g_2$ then we need to apply g_1 twice followed by a single application of g_2.)

4.3 Performance Details

As remarked in the introduction, we implemented the above variant of evaluating AES homomorphically on a very large memory machine; namely a machine with 256 GB of RAM. Firstly parameters were selected, as in Appendix C to cope with 60 levels of computation, and a public/private key pair was generated; along with the key-switching data for multiplication operations and conjugation with-respect-to the Galois group.

As input to the actual computation was an AES plaintext block and the eleven round keys; each of which was encrypted using our homomorphic encryption scheme. Thus the input consisted of eleven packed ciphertexts. Producing the encrypted key schedule took around half an hour. To evaluate the entire ten rounds of AES took about 47 hours (or under two days); however each of our ciphertexts could hold 864 plaintext slots of elements in $\mathbb{F}_{2^8}$, thus we could have processed 54 such AES blocks in this time period. This would result in a throughput of just under one hour per AES block.

We note that as the algorithm progressed the operations became faster. The first round of the AES function took 10 hours, whereas the penultimate round took 3 hours and the last round took 45 minutes. Recall, the last AES round is somewhat simpler as it does not involve a MixColumns operation.

Whilst our other two implementation choices (given in Appendix D) may seem to appear to be more efficient, the increase in memory requirements and data actually makes them less attractive when encrypting a single block. For example just encrypting the key schedule in the Byte-Sliced variant takes 6 hours (with 50 levels), with an entire round taking 122 hours (28 hours for the first round, 8 for both the penultimate and final rounds). This however equates to an amortized time of just over 10 minutes per block.

The Bit-Sliced variant requires over 150 hours to just encrypt the key schedule (with 60 levels), and evaluating a single round takes so long that our program is timed out before even a single round is evaluated.

References

[1] Benny Applebaum, David Cash, Chris Peikert, and Amit Sahai. Fast cryptographic primitives and circular-secure encryption based on hard learning problems. In *CRYPTO*, volume 5677 of *Lecture Notes in Computer Science*, pages 595–618. Springer, 2009.

[2] Sanjeev Arora and Rong Ge. New algorithms for learning in the presence of errors. Manuscript, 2011.

[3] Joan Boyar and René Peralta. A depth-16 circuit for the AES S-box. Manuscript, 2011.

[4] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. Fully homomorphic encryption without bootstrapping. Manuscript at http://eprint.iacr.org/2011/277, 2011.

[5] Zvika Brakerski and Vinod Vaikuntanathan. Efficient fully homomorphic encryption from (standard) LWE, 2011.

[6] Zvika Brakerski and Vinod Vaikuntanathan. Fully homomorphic encryption from ring-LWE and security for key dependent messages. In *Advances in Cryptology - CRYPTO 2011*, volume 6841 of *Lecture Notes in Computer Science*, pages 505–524. Springer, 2011.

[7] Jean-Sébastien Coron, Avradip Mandal, David Naccache, and Mehdi Tibouchi. Fully-homomorphic encryption over the integers with shorter public-keys. Manuscript, to appear in Crypto 2011.

[8] Ivan Damgård and Marcel Keller. Secure multiparty aes. In *Proc. of Financial Cryptography 2010*, volume 6052 of *LNCS*, pages 367–374, 2010.

[9] Ivan Damgård, Valerio Pastro, Nigel P. Smart, and Sarah Zakarias. Multiparty computation from somewhat homomorphic encryption. Manuscript, 2011.

[10] Nicolas Gama and Phong Q. Nguyen. Predicting lattice reduction. In *EUROCRYPT*, volume 4965 of *Lecture Notes in Computer Science*, pages 31–51. Springer, 2008.

[11] Craig Gentry. Fully homomorphic encryption using ideal lattices. In Michael Mitzenmacher, editor, *STOC*, pages 169–178. ACM, 2009.

[12] Craig Gentry and Shai Halevi. Implementing gentry's fully-homomorphic encryption scheme. In *EUROCRYPT*, volume 6632 of *Lecture Notes in Computer Science*, pages 129–148. Springer, 2011.

[13] Craig Gentry, Shai Halevi, and Nigel Smart. Fully homomorphic encryption with polylog overhead. Manuscript, 2011.

[14] Yan Huang, David Evans, Jonathan Katz, and Lior Malka. Faster secure two-party computation using garbled circuits. In *USENIX Security Symposium*, 2011.

[15] C. Orlandi J.B. Nielsen, P.S. Nordholt and S. Sheshank. A new approach to practical active-secure two-party computation. Manuscript, 2011.

[16] Kristin Lauter, Michael Naehrig, and Vinod Vaikuntanathan. Can homomorphic encryption be practical? Manuscript at http://www.codeproject.com/News/15443/Can-Homomorphic-Encryption-be-Practical.aspx, 2011.

[17] Richard Lindner and Chris Peikert. Better key sizes (and attacks) for lwe-based encryption. In *CT-RSA*, volume 6558 of *Lecture Notes in Computer Science*, pages 319–339. Springer, 2011.

[18] Daniele Micciancio and Oded Regev. *Lattice-based cryptography*, pages 147–192. Springer, 2009.

[19] Benny Pinkas, Thomas Schneider, Nigel P. Smart, and Steven C. Williams. Secure two-party computation is practical. In *Proc. ASIACRYPT 2009*, volume 5912 of *LNCS*, pages 250–267, 2009.

[20] Nigel P. Smart and Frederik Vercauteren. Fully homomorphic encryption with relatively small key and ciphertext sizes. In *Public Key Cryptography - PKC'10*, volume 6056 of *Lecture Notes in Computer Science*, pages 420–443. Springer, 2010.

[21] Nigel P. Smart and Frederik Vercauteren. Fully homomorphic SIMD operations. Manuscript at http://eprint.iacr.org/2011/133, 2011.

A More Details

Following [4, 13, 21] we define utilize rings defined by cyclotomic polynomials, $\mathbb{A} = \mathbb{Z}[X]/\Phi_m(X)$. We let $\mathbb{A}_q$ denote the set of elements of this ring reduced modulo various (possibly composite) moduli q. The ring $\mathbb{A}$ is the ring of integers of a the mth cyclotomic number field K.

A.1 Plaintext Slots

In our scheme plaintexts will be elements of $\mathbb{A}_2$, and the polynomial $\Phi_m(X)$ factors modulo 2 into ℓ irreducible factors, $\Phi_m(X) = F_1(X) \cdot F_2(X) \cdots F_\ell(X) \pmod 2$, all of degree $d = \phi(m)/\ell$. Just as in [4, 13, 21] each factor corresponds to a "plaintext slot". That is, we view a polynomial $a \in \mathbb{A}_2$ as representing an ℓ-vector $(a \bmod F_i)_{i=1}^{\ell}$.

It is standard fact that the Galois group $\mathcal{G}\mathsf{al} = \mathcal{G}\mathsf{al}(\mathbb{Q}(\zeta_m)/\mathbb{Q})$ consists of the mappings $\kappa_k : a(X) \mapsto a(x^k) \bmod \Phi_m(X)$ for all k co-prime with m, and that it is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^*$. As noted in [13], for each $i, j \in \{1, 2, \ldots, \ell\}$ there is an element $\kappa_k \in \mathcal{G}\mathsf{al}$ which sends an element in slot i to an element in slot j. Namely, if $b = \kappa_i(a)$ then the element in the j'th slot of b is the same as that in the i'th slot of a. In addition $\mathcal{G}\mathsf{al}$ contains the Frobenius elements, $X \longrightarrow X^{2^i}$, which also act as Frobenius on the individual slots separately.

We will be specifically interested in arithmetic in a specific finite field $\mathbb{F}_{2^n}$ where n divides d, given by a specific polynomial representation $\mathbb{F}_2[X]/G(X)$ (with G of degree n). Specifically we use the polynomial defining the AES operations, $G(X) = X^8 + X^4 + X^3 + X + 1$. Since n divides d, each plaintext slot contains a copy of $\mathbb{F}_{2^n}$ and so we can think of the plaintext space as containing ℓ copies of $\mathbb{F}_{2^n}$.

A.2 Canonical Embedding Norm

Following [13], we use as the "size" of a polynomial $a \in \mathbb{A}$ the l_∞ norm of its canonical embedding. Recall that the canonical embedding of $a \in \mathbb{A}$ into $\mathbb{C}^{\phi(m)}$ is the $\phi(m)$-vector of complex numbers $\sigma(a) = (a(\zeta_m^i))_i$ where ζ_m is a complex primitive m-th root of unity and the indexes i range over all of $(\mathbb{Z}/m\mathbb{Z})^*$. We call the norm of $\sigma(a)$ the *canonical embedding norm* of a, and denote it by

$$\|a\|_\infty^{\mathsf{can}} = \|\sigma(a)\|_\infty.$$

We will make use of the following properties of $\|\cdot\|_\infty^{\mathsf{can}}$:

- For all $a, b \in \mathbb{A}$ we have $\|a \cdot b\|_\infty^{\mathsf{can}} \leq \|a\|_\infty^{\mathsf{can}} \cdot \|b\|_\infty^{\mathsf{can}}$.

- For all $a \in \mathbb{A}$ we have $\|a\|_\infty^{\mathsf{can}} \leq \|a\|_1$.
- There is a ring constant c_m (depending only on m) such that $\|a\|_\infty \leq c_m \cdot \|a\|_\infty^{\mathsf{can}}$ for all $a \in \mathbb{A}$.

The ring constant c_m is defined by $c_m = \|\mathsf{CRT}_m^{-1}\|_\infty$ where CRT_m is the CRT matrix for m, i.e. the Vandermonde matrix over the complex primitive m-th roots of unity. Asymptotically the value c_m can grow super-polynomially with m, but for the "small" values of m one would use in practice values of c_m can be evaluated directly. See [9] for a discussion of c_m.

Canonical Reduction. When working with elements in $\mathbb{A}_q$ for some integer modulus q, we sometimes need a version of the canonical embedding norm that plays nice with reduction modulo q. Following [13], we define the *canonical embedding norm reduced modulo* q of an element $a \in \mathbb{A}$ as the smallest canonical embedding norm of any a' which is congruent to a modulo q. We denote it as

$$|a|_q^{can} \stackrel{\text{def}}{=} \min\{\, \|a'\|_\infty^{\mathsf{can}} : \; a' \in \mathbb{A},\; a' \equiv a \pmod q \,\}.$$

We sometimes also denote the polynomial where the minimum is obtained by $[a]_q^{\mathsf{can}}$, and call it the *canonical reduction* of a modulo q. Neither the canonical embedding norm nor the canonical reduction is used in the scheme itself, it is only in the analysis of it that we will need them. We note that (trivially) we have $|a|_q^{\mathsf{can}} \leq \|a\|_\infty^{\mathsf{can}}$.

A.3 Double CRT Representation

As noted in Section 2, we usually represent an element $a \in A_q$ via double-CRT representation, with respect to both the polynomial factor of $\Phi_m(X)$ and the integer factors of q. Specifically, we assume that $\mathbb{Z}/q\mathbb{Z}$ contains a primitive m-th root of unity (call it ζ), so $\Phi_m(X)$ factors modulo q to linear terms $\Phi_m(X) = \prod_{i\in(\mathbb{Z}/m\mathbb{Z})^*}(X - \zeta^j) \pmod q$. We also denote q's prime factorization by $q = \prod_{i=0}^{t} p_i$. Then a polynomial $a \in \mathbb{A}_q$ is represented as the $(t+1) \times \phi(m)$ matrix of its evaluation at the roots of $\Phi_m(X)$ modulo p_i for $i = 0, \ldots, t$:

$$\mathsf{dble\text{-}CRT}^t(a) = \left(a\left(\zeta^j\right) \bmod p_i\right)_{0\leq i\leq t, j\in(\mathbb{Z}/m\mathbb{Z})^*}.$$

The double CRT representation can be computed using $t+1$ invocations of the FFT algorithm modulo the p_i, picking only the FFT coefficients which correspond to elements in $(\mathbb{Z}/m\mathbb{Z})^*$. To invert this representation we invoke the inverse FFT algorithm $t+1$ times on a vector of length m consisting of the thinned out values padded with zeros, then apply the Chinese Remainder Theorem, and then reduce modulo $\Phi_m(X)$ and q.

Addition and multiplication in $\mathbb{A}_q$ can be computed as component-wise addition and multiplication of the entries in the two tables (modulo the appropriate primes p_i),

$$\begin{aligned}\mathsf{dble\text{-}CRT}^t(a+b) &= \mathsf{dble\text{-}CRT}^t(a) + \mathsf{dble\text{-}CRT}^t(b)\\ \mathsf{dble\text{-}CRT}^t(a\cdot b) &= \mathsf{dble\text{-}CRT}^t(a) \cdot \mathsf{dble\text{-}CRT}^t(b).\end{aligned}$$

Also, for an element of the Galois group $\kappa_k \in \mathcal{G}\mathsf{al}$ (which maps $a(X) \in \mathbb{A}$ to $a(X^k) \bmod \Phi_m(X)$), we can evaluate $\kappa_k(a)$ on the double-CRT representation of a just by permuting the columns in the matrix, sending each column j to column $j \cdot k \bmod m$.

A.4 Sampling From $\mathbb{A}_q$

At various points we will need to sample from $\mathbb{A}_q$ with different distributions, as described below. We denote choosing the element $a \in \mathbb{A}$ according to distribution $\mathcal{D}$ by $a \leftarrow \mathcal{D}$. The distributions below are described as over $\phi(m)$-vectors, but we always consider them as distributions over the ring $\mathbb{A}$, by identifying a polynomial $a \in \mathbb{A}$ with its coefficient vector.

The uniform distribution $\mathcal{U}_q$: This is just the uniform distribution over $(\mathbb{Z}/q\mathbb{Z})^{\phi(m)}$, which we identify with $(\mathbb{Z} \cap (-q/2, q/2])^{\phi(m)})$. Note that it is easy to sample from $\mathcal{U}_q$ directly in double-CRT representation.

The "discrete Gaussian" $\mathcal{DG}_q(\sigma^2)$: Let $\mathcal{N}(0, \sigma^2)$ denote the normal (Gaussian) distribution on real numbers with zero-mean and variance σ^2, we use drawing from $\mathcal{N}(0, \sigma^2)$ and rounding to the nearest integer as an approximation to the discrete Gaussian distribution. Namely, the distribution $\mathcal{DG}_{q_t}(\sigma^2)$ draws a real ϕ-vector according to $\mathcal{N}(0, \sigma^2)^{\phi(m)}$, rounds it to the nearest integer vector, and outputs that integer vector reduced modulo q (into the interval $(-q/2, q/2]$).

Sampling small polynomials, $\mathcal{ZO}(p)$ and $\mathcal{HWT}(h)$: These distributions produce vectors in $\{0, \pm 1\}^{\phi(m)}$.

For a real parameter $\rho \in [0, 1]$, $\mathcal{ZO}(p)$ draws each entry in the vector from $\{0, \pm 1\}$, with probability $\rho/2$ for each of -1 and $+1$, and probability of being zero $1 - \rho$.

For an integer parameter $h \leq \phi(m)$, the distribution $\mathcal{HWT}(h)$ chooses a vector uniformly at random from $\{0, \pm 1\}^{\phi(m)}$, subject to the conditions that it has exactly h nonzero entries.

A.5 Canonical embedding norm of random polynomials

In the coming sections we will need to bound the canonical embedding norm of polynomials that are produced by the distributions above, as well as products of such polynomials. In some cases it is possible to analyze the norm rigorously using Chernoff and Hoeffding bounds, but to set the parameters of our scheme we instead use a heuristic approach that yields better constants:

Let $a \in \mathbb{A}$ be a polynomial that was chosen by one of the distributions above, hence all the (nonzero) coefficients in a are IID (independently identically distributed). For a complex primitive m-th root of unity ζ_m, the evaluation $a(\zeta_m)$ is the inner product between the coefficient vector of a and the fixed vector $\mathbf{z}_m = (1, \zeta_m, \zeta_m^2, \ldots)$, which has Euclidean norm exactly $\sqrt{\phi(m)}$. Hence the random variable $a(\zeta_m)$ has variance $V = \sigma^2\phi(m)$, where σ^2 is the variance of each coefficient of a. Specifically, when $a \leftarrow \mathcal{U}_q$ then each coefficient has variance $q^2/12$, so we get variance $V_U = q^2\phi(m)/12$. When $a \leftarrow \mathcal{DG}_q(\sigma^2)$ we get variance $V_G \approx \sigma^2\phi(m)$, and when $a \leftarrow \mathcal{ZO}(\rho)$ we get variance $V_Z = \rho\phi(m)$. When choosing $a \leftarrow \mathcal{HWT}(h)$ we get a variance of $V_H = h$ (but not $\phi(m)$, since a has only h nonzero coefficients).

Moreover, the random variable $a(\zeta_m)$ is a sum of many IID random variables, hence by the law of large numbers it is distributed similarly to a complex Gaussian random variable of the specified variance.[6] We therefore use $6\sqrt{V}$ (i.e. six standard deviations) as a high-probability bound on the size of $a(\zeta_m)$. Since the evaluation of a at all the roots of unity obeys the same bound, we use six standard deviations as our bound on the canonical embedding norm of a. (We chose six standard deviations since $\mathrm{erfc}(6) \approx 2^{-55}$, which is good enough for us even when using the union bound and multiplying it by $\phi(m) \approx 2^{16}$.)

In many cases we need to bound the canonical embedding norm of a product of two such "random polynomials". In this case our task is to bound the magnitude of the product of two random variables, both are distributed close to Gaussians, with variances σ_a^2, σ_b^2, respectively. For this case we use $16\sigma_a\sigma_b$ as our

[6] The mean of $a(\zeta_m)$ is zero, since the coefficients of a are chosen from a zero-mean distribution.

bound, since $\text{erfc}(4) \approx 2^{-25}$, so the probability that both variables exceed their standard deviation by more than a factor of four is roughly 2^{-50}.

B The Basic Scheme

We now define our leveled HE scheme on L levels; including the Modulus-Switching and Key-Switching operations and the procedures for KeyGen, Enc, Dec, and for Add, Mult, Scalar-Mult, and Automorphism.

Recall that a ciphertext vector $\mathbf{c}$ in the cryptosystem is a valid encryption of $a \in \mathbb{A}$ with respect to secret key $\mathbf{s}$ and modulus q if $[[\langle \mathbf{c}, \mathbf{s} \rangle]_q]_2 = a$, where the inner product is over $\mathbb{A} = \mathbb{Z}[X]/\Phi_m(X)$, the operation $[\cdot]_q$ denotes modular reduction in coefficient representation into the interval $(-q/2, +q/2]$, and we require that the "noise" $[\langle \mathbf{c}, \mathbf{s} \rangle]_q$ is sufficiently small (in canonical embedding norm reduced mod q). In our implementation a "normal" ciphertext is a 2-vector $\mathbf{c} = (c_0, c_1)$, and a "normal" secret key is of the form $\mathbf{s} = (1, -\mathfrak{s})$, hence decryption takes the form

$$a \leftarrow [c_0 - c_1 \cdot \mathfrak{s}]_q \bmod 2. \tag{2}$$

B.1 Our Moduli Chain

We define the chain of moduli for our depth-L homomorphic evaluation by choosing L "small primes" $p_0, p_1, \ldots, p_{L-1}$ and the t'th modulus in our chain is defined as $q_t = \prod_{j=0}^{t} p_j$. (The sizes will be determined later.) The primes p_i's are chosen so that for all i, $\mathbb{Z}/p_i\mathbb{Z}$ contains a primitive m-th root of unity. Hence we can use our double-CRT representation for all $\mathbb{A}_{q_t}$.

This choice of moduli makes it easy to get a level-$(t-1)$ representation of $a \in \mathbb{A}$ from its level-t representation. Specifically, given the level-t double-CRT representation $\mathsf{dble\text{-}CRT}^t(a)$ for some $a \in \mathbb{A}_{q_t}$, we can simply remove from the matrix the row corresponding to the last small prime p_t, thus obtaining a level-$(t-1)$ representation of $a \bmod q_{t-1} \in \mathbb{A}_{q_{t-1}}$. Similarly we can get the double-CRT representation for lower levels by removing more rows. By a slight abuse of notation we write $\mathsf{dble\text{-}CRT}^{t'}(a) = \mathsf{dble\text{-}CRT}^t(a) \bmod q_{t'}$ for $t' < t$.

Recall that encryption produces ciphertext vectors valid with respect to the largest modulus q_{L-1} in our chain, and we obtain ciphertext vectors valid with respect to smaller moduli whenever we apply modulus-switching to decrease the noise magnitude. As described in Section 3.3, our implementation *dynamically* adjust levels, performing modulus switching when the dynamically-computed noise estimate becomes too large. Hence each ciphertext in our scheme is tagged with both its level t (pinpointing the modulus q_t relative to which this ciphertext is valid), and an estimate ν on the noise magnitude in this ciphertext. In other words, a ciphertext is a triple $(\mathbf{c}, t, \nu)$ with $0 \le t \le L-1$, $\mathbf{c}$ a vector over $\mathbb{A}_{q_t}$, and ν a real number which is used as our noise estimate.

B.2 Modulus Switching

The operation $\mathsf{SwitchModulus}(\mathfrak{c})$ takes the ciphertext $\mathfrak{c} = ((c_0, c_1), t, \nu)$ defined modulo q_t and produces a ciphertext $\mathfrak{c}' = ((c'_0, c'_1), t-1, \nu')$ defined modulo q_{t-1}, Such that $[c_0 - \mathfrak{s} \cdot c_1]_{q_t} \equiv [c'_0 - \mathfrak{s} \cdot c'_1]_{q_{t-1}} \pmod 2$, and ν' is smaller than ν. This procedure makes use of the function $\mathsf{Scale}(x, q, q')$ that takes an element $x \in \mathbb{A}_q$ and returns an element $y \in \mathbb{A}_{q'}$ such that in coefficient representation it holds that $y \equiv x \pmod 2$, and y is the closest element to $(q'/q) \cdot x$ that satisfies this mod-2 condition.

To maintain the noise estimate, the procedure uses the pre-set ring-constant c_m (cf. Appendix A.2) and also a pre-set constant B_{scale} which is meant to bound the magnitude of the added noise term from this operation. It works as follows:

SwitchModulus$((c_0, c_1), t, \nu)$:

1. If $t < 1$ then abort; // Sanity check
2. $\nu' \leftarrow \frac{q_{t-1}}{q_t} \cdot \nu + B_{\mathsf{scale}}$; // Scale down the noise estimate
3. If $\nu' > q_{t-1}/2c_m$ then abort; // Another sanity check
4. $c'_i \leftarrow \mathsf{Scale}(c_i, q_t, q_{t-1})$ for $i = 0, 1$; // Scale down the vector
5. Output $((c'_0, c'_1), t-1, \nu')$.

The constant B_{scale} is set as $B_{\mathsf{scale}} = 2\sqrt{\phi(m)/3} \cdot (8\sqrt{h} + 3)$, where h is the Hamming weight of the secret key. (In our implementation we use $h = 64$, so we get $B_{\mathsf{scale}} \approx 77\sqrt{\phi(m)}$.) To justify this choice, we apply to the proof of the modulus switching lemma from [13, Lemma 13] (in the full version), relative to the canonical embedding norm. In that proof it is shown that when the noise magnitude in the input ciphertext $\mathbf{c} = (c_0, c_1)$ is bounded by ν, then the noise magnitude in the output vector $\mathbf{c}' = (c'_0, c'_1)$ is bounded by $\nu' = \frac{q_{t-1}}{q_t} \cdot \nu + \| \langle \mathbf{s}, \tau \rangle \|^{\mathsf{can}}_{\infty}$, provided that the last quantity is smaller than $q_{t-1}/2$.

Above τ is the "rounding error" vector, namely $\tau \stackrel{\text{def}}{=} (\tau_0, \tau_1) = (c'_0, c'_1) - \frac{q_{t-1}}{q_t}(c_0, c_1)$. Heuristically assuming that τ behaves as if its coefficients are chosen uniformly in $[-1, +1]$, the evaluation $\tau_i(\zeta)$ at an m-th root of unity ζ_m is distributed close to a Gaussian complex with variance $\phi(m)/3$. Also, $\mathfrak{s}$ was drawn from $\mathcal{HWT}(h)$ so $\mathfrak{s}(\zeta_m)$ is distributed close to a Gaussian complex with variance h. Hence we expect $\tau_1(\zeta)\mathfrak{s}(\zeta)$ to have magnitude at most $16\sqrt{\phi(m)/3 \cdot h}$ (recall that we use $h = 64$). We can similarly bound $\tau_0(\zeta_m)$ by $6\sqrt{\phi(m)/3}$, and therefore the evaluation of $\langle \mathbf{s}, \tau \rangle$ at ζ_m is bounded in magnitude (whp) by:

$$16\sqrt{\phi(m)/3 \cdot h} + 6\sqrt{\phi(m)/3} \;=\; 2\sqrt{\phi(m)/3} \cdot \left(8\sqrt{h} + 3\right) \;\approx\; 77\sqrt{\phi(m)} \;=\; B_{\mathsf{scale}} \tag{3}$$

B.3 Key Switching

After some homomorphic evaluation operations we have on our hands not a "normal" ciphertext which is valid relative to "normal" secret key, but rather an "extended ciphertext" $((d_0, d_1, d_2), q_t, \nu)$ which is valid with respect to an "extended secret key" $\mathbf{s}' = (1, -\mathfrak{s}, -\mathfrak{s}')$. Namely, this ciphertext encrypts the plaintext $a \in \mathbb{A}$ via

$$a \;=\; \left[\left[d_0 - \mathfrak{s} \cdot d_1 - \mathfrak{s}' \cdot d_2\right]_{q_t}\right]_2$$

and the magnitude of the noise $\left[d_0 - \mathfrak{s} \cdot d_1 - d_2 \cdot \mathfrak{s}'\right]_{q_t}$ is bounded by ν. In our implementation, the component $\mathfrak{s}$ is always the same element $\mathfrak{s} \in \mathbb{A}$ that was drawn from $\mathcal{HWT}(h)$ during key generation, but $\mathfrak{s}'$ can vary depending on the operation. (See the description of multiplication and automorphisms below.)

To enable that translation, we use some "key switching matrices" that are included in the public key. (In our implementation these "matrices" have dimension 2×1, i.e., the consist of only two elements from $\mathbb{A}$.) As explained in Section 3.1, we save on space and time by artificially "boosting" the modulus we use from q_t up to $P \cdot q_t$ for some "large" modulus P. We note that in order to represent elements in $\mathbb{A}_{Pq_t}$ using our dble-CRT representation we need to choose P so that $\mathbb{Z}/P\mathbb{Z}$ also has primitive m-th roots of unity. (In fact in our implementation we pick P to be a prime.)

The key-switching "matrix". Denote by $Q = P \cdot q_{L-2}$ the largest modulus relative to which we need to generate key-switching matrices. To generate the key-switching matrix from $\mathbf{s}' = (1, -\mathfrak{s}, -\mathfrak{s}')$ to $\mathbf{s} = (1, -\mathfrak{s})$ (note that both keys share the same element $\mathfrak{s}$), we choose two element, one uniform and the other from our "discrete Gaussian",

$$a_{\mathfrak{s},\mathfrak{s}'} \leftarrow \mathcal{U}_Q \;\text{ and }\; e_{\mathfrak{s},\mathfrak{s}'} \leftarrow \mathcal{DG}_Q(\sigma^2),$$

where the variance σ is a global parameter (that we later set as $\sigma = 3.2$). The "key switching matrix" then consists of the single column vector

$$W[\mathfrak{s}' \to \mathfrak{s}] \;=\; \begin{pmatrix} b_{\mathfrak{s},\mathfrak{s}'} \\ a_{\mathfrak{s},\mathfrak{s}'} \end{pmatrix}, \text{ where } b_{\mathfrak{s},\mathfrak{s}'} \stackrel{\text{def}}{=} \left[\mathfrak{s} \cdot a_{\mathfrak{s},\mathfrak{s}'} + 2e_{\mathfrak{s},\mathfrak{s}'} + P\mathfrak{s}'\right]_Q. \tag{4}$$

Note that W above is defined modulo $Q = Pq_{L-2}$, but we need to use it relative to $Q_t = Pq_t$ for whatever the current level t is. Hence before applying the key switching procedure at level t, we reduce W modulo Q_t to get $W_t \stackrel{\text{def}}{=} [W]_{Q_t}$. It is important to note that since Q_t divides Q then W_t is indeed a key-switching matrix. Namely it is of the form $(b, a)^T$ with $a \in \mathcal{U}_{Q_t}$ and $b = [\mathfrak{s} \cdot a + 2e_{\mathfrak{s},\mathfrak{s}'} + P\mathfrak{s}']_{Q_t}$ (with respect to the same element $e_{\mathfrak{s},\mathfrak{s}'} \in \mathbb{A}$ from above).

The SwitchKey **procedure.** Given the extended ciphertext $\mathfrak{c} = ((d_0, d_1, d_2), t, \nu)$ and the key-switching matrix $W_t = (b, a)^T$, the procedure $\mathsf{SwitchKey}_{W_t}(\mathfrak{c})$ proceeds as follows:[7]

$\underline{\mathsf{SwitchKey}_{(b,a)}((d_0, d_1, d_2), t, \nu)}$:

1. Set $\begin{pmatrix} c'_0 \\ c'_1 \end{pmatrix} \leftarrow \left[\begin{pmatrix} Pd_0 & b \\ Pd_1 & a \end{pmatrix}\begin{pmatrix} 1 \\ d_2 \end{pmatrix}\right]_{Q_t}$; // The actual key-switching operation
2. $c''_i \leftarrow \mathsf{Scale}(c'_i, Q_t, q_t)$ for $i = 0, 1$; // Scale the vector back down to q_t
3. $\nu' \leftarrow \nu + B_{\mathsf{Ks}} \cdot q_t/P + B_{\mathsf{scale}}$; // The constant B_{Ks} is determined below
4. Output $((c''_0, c''_1), t, \nu')$.

To argue correctness, observe that although the "actual key switching operation" from above looks superficially different from the standard key-switching operation $\mathbf{c}' \leftarrow W \cdot \mathbf{c}$, it is merely an optimization that takes advantage of the fact that both vectors $\mathbf{s}'$ and $\mathbf{s}$ share the element $\mathfrak{s}$. Indeed, we have the equality over $\mathbb{A}_{Q_t}$:

$$\begin{aligned} c'_0 - \mathfrak{s} \cdot c'_1 &= [(P \cdot d_0) + d_2 \cdot b_{\mathfrak{s},\mathfrak{s}'} - \mathfrak{s} \cdot \big((P \cdot d_1) + d_2 \cdot a_{\mathfrak{s},\mathfrak{s}'}\big) \\ &= P \cdot (d_0 - \mathfrak{s} \cdot d_1 - \mathfrak{s}' d_2) + 2 \cdot d_2 \cdot \epsilon_{\mathfrak{s},\mathfrak{s}'}, \end{aligned}$$

so as long as both sides are smaller than Q_t we have the same equality also over $\mathbb{A}$ (without the mod-Q_t reduction), which means that we get

$$[c'_0 - \mathfrak{s} \cdot c'_1]_{Q_t} \;=\; [P \cdot (d_0 - \mathfrak{s} \cdot d_1 - \mathfrak{s}' d_2) + 2 \cdot d_2 \cdot \epsilon_{\mathfrak{s},\mathfrak{s}'}]_{Q_t} \;\equiv\; [d_0 - \mathfrak{s} \cdot d_1 - \mathfrak{s}' d_2]_{Q_t} \pmod 2.$$

To analyze the size of the added term $2d_2\epsilon_{\mathfrak{s},\mathfrak{s}'}$, we can assume heuristically that d_2 behaves like a uniform polynomial drawn from $\mathcal{U}_{q_t}$, hence $d_2(\zeta_m)$ for a complex root of unity ζ_m is distributed close to a complex Gaussian with variance $q_t^2\phi(m)/12$. Similarly $\epsilon_{\mathfrak{s},\mathfrak{s}'}(\zeta_m)$ is distributed close to a complex Gaussian with variance $\sigma^2\phi(m)$, so $2d_2(\zeta)\epsilon(\zeta)$ can be modeled as a product of two Gaussians, and we expect that with overwhelming probability it remains smaller than $2 \cdot 16 \cdot \sqrt{q_t^2\phi(m)/12 \cdot \sigma^2\phi(m)} \;=\; \frac{16}{\sqrt{3}} \cdot \sigma q_t \phi(m)$. This yields a heuristic bound $16/\sqrt{3} \cdot \sigma\phi(m) \cdot q_t \;=\; B_{\mathsf{Ks}} \cdot q_t$ on the canonical embedding norm of the added noise term, and if the total noise magnitude does not exceed $Q_t/2c_m$ then also in coefficient representation everything remains below $Q_t/2$. Thus our constant B_{Ks} is set as

$$\frac{16\sigma\phi(m)}{\sqrt{3}} \;\approx\; 9\sigma\phi(m) \;=\; B_{\mathsf{Ks}} \tag{5}$$

[7] For simplicity we describe the SwitchKey procedure as if it always switches back to mod-q_t, but in reality if the noise estimate is large enough then it can switch directly to q_{t-1} instead.

Finally, dividing by P (which is the effect of the Scale operation), we obtain the final ciphertext that we require, and the noise magnitude is divided by P (except for the added B_{scale} term).

B.4 Key-Generation, Encryption, and Decryption

The procedures below depend on many parameters, h, σ, m, the primes p_i and P, etc. These parameters will be determined later.

KeyGen(): Given the parameters, the key generation procedure chooses a low-weight secret key and then generates an LWE instance relative to that secret key. Namely, we choose

$$\mathfrak{s} \leftarrow \mathcal{HWT}(h), \quad a \leftarrow \mathcal{U}_{q_{L-1}}, \quad \text{and } e \leftarrow \mathcal{DG}_{q_{L-1}}(\sigma^2)$$

Then sets the secret key as $\mathfrak{s}$ and the public key as (a, b) where $b = [a \cdot s + 2e]_{q_{L-1}}$.

In addition, the key generation procedure adds to the public key some key-switching "matrices", as described in Appendix B.3. Specifically the matrix $W[\mathfrak{s}^2 \to \mathfrak{s}]$ for use in multiplication, and some matrices $W[\kappa_i(\mathfrak{s}) \to \mathfrak{s}]$ for use in automorphisms, for $\kappa_i \in \mathcal{G}\mathsf{al}$ whose indexes generates $(\mathbb{Z}/m\mathbb{Z})^*$ (including in particular κ_2).

$\mathsf{Enc}_{\mathfrak{pk}}(\mathbf{m})$: To encrypt an element $m \in \mathbb{A}_2$, we choose one "small polynomial" (with $0, \pm 1$ coefficients) and two Gaussian polynomials (with variance σ^2),

$$v \leftarrow \mathcal{ZO}(0.5) \quad \text{and} \quad e_0, e_1 \leftarrow \mathcal{DG}_{q_{L-1}}(\sigma^2)$$

Then we set $c_0 = b \cdot v + 2 \cdot e_0 + m$, $c_1 = a \cdot v + 2 \cdot e_1$, and set the initial ciphertext as $\mathfrak{c}' = (c_0, c_1, L-1, B_{\mathsf{clean}})$, where B_{clean} is a parameter that we determine below.

The noise magnitude in this ciphertext (B_{clean}) is a little larger than what we would like, so before we start computing on it we do one modulus-switch. That is, the encryption procedure sets $\mathfrak{c} \leftarrow \mathsf{SwitchModulus}(\mathfrak{c}')$ and outputs $\mathfrak{c}$. We can deduce a value for B_{clean} as follows:

$$\begin{aligned}
\left|c_0 - \mathfrak{s} \cdot c_1\right|_{q_t}^{\mathsf{can}} &\le \|c_0 - \mathfrak{s} \cdot c_1\|_\infty^{\mathsf{can}} \\
&= \|((a \cdot s + 2 \cdot e) \cdot v + 2 \cdot e_0 + \mathbf{m} - (a \cdot v + 2 \cdot e_1) \cdot \mathfrak{s}\|_\infty^{\mathsf{can}} \\
&= \|\mathbf{m} + 2 \cdot (e \cdot v + e_0 - e_1 \cdot \mathfrak{s})\|_\infty^{\mathsf{can}} \\
&\le \|\mathbf{m}\|_\infty^{\mathsf{can}} + 2 \cdot (\|e \cdot v\|_\infty^{\mathsf{can}} + \|e_0\|_\infty^{\mathsf{can}} + \|e_1 \cdot \mathfrak{s}\|_\infty^{\mathsf{can}})
\end{aligned}$$

Using our complex Gaussian heuristic from Appendix A.5, we can bound the canonical embedding norm of the randomized terms above by

$$\|e \cdot v\|_\infty^{\mathsf{can}} \ \le \ 16\sqrt{2} \cdot \sigma\phi(m), \quad \|e_0\|_\infty^{\mathsf{can}} \ \le \ 6\sigma\sqrt{\phi(m)}, \quad \|e_1 \cdot \mathfrak{s}\|_\infty^{\mathsf{can}} \ \le \ 16\sigma\sqrt{h \cdot \phi(m)}$$

Also, the norm of the input message m is clearly bounded by $\phi(m)$, hence (when we substitute our parameters $h = 64$ and $\sigma = 3.2$) we get the bound

$$\phi(m) + 32\sqrt{2} \cdot \sigma\phi(m) + 12\sigma\sqrt{\phi(m)} + 32\sigma\sqrt{h \cdot \phi(m)} \ \approx \ 146\phi(m) + 858\sqrt{\phi(m)} \ = \ B_{\mathsf{clean}} \qquad (6)$$

Our goal in the initial modulus switching from q_{L-1} to q_{L-2} is to reduce the noise from its initial level of $B_{\mathsf{clean}} = \Theta(\phi(m))$ to our base-line bound of $B = \Theta(\sqrt{\phi(m)})$ which is determined in Equation (12) below.

$\underline{\mathsf{Dec}_{\mathfrak{pt}}(\mathfrak{c})}$: Decryption of a ciphertext (c_0, c_1, t, ν) at level t is performed by setting $m' \leftarrow [c_0 - \mathfrak{s} \cdot c_1]_{q_t}$, then converting m' to coefficient representation and outputting $m' \bmod 2$. This procedure works when $c_m \cdot \nu < q_t/2$, so this procedure only applies when the constant c_m for the field $\mathbb{A}$ is known and relatively small (which as we mentioned above will be true for all practical parameters). Also, we must pick the smallest prime $q_0 = p_0$ large enough, as described in Appendix C.2.

B.5 Homomorphic Operations

$\underline{\mathsf{Add}(\mathfrak{c}, \mathfrak{c}')}$: Given two ciphertexts $\mathfrak{c} = ((c_0, c_1), t, \nu)$ and $\mathfrak{c}' = ((c'_0, c'_1), t', \nu')$, representing messages $\mathbf{m}, \mathbf{m}' \in \mathbb{A}_2$, this algorithm forms a ciphertext $\mathfrak{c}_a = ((a_0, a_1), t_a, \nu_a)$ which encrypts the message $\mathbf{m}_a = \mathbf{m} + \mathbf{m}'$.

If the two ciphertexts do not belong to the same level then we reduce the larger one modulo the smaller of the two moduli, thus bringing them to the same level. (This simple modular reduction works as long as the noise magnitude is smaller than the smaller of the two moduli, if this condition does not hold then we need to do modulus switching rather than simple modular reduction.) Once the two ciphertexts are at the same level (call it t''), we just add the two ciphertext vectors and two noise estimates to get

$$\mathfrak{c}_a \;=\; \left(\left([c_0 + c'_0]_{q_{t''}}, [c_1 + c'_1]_{q_{t''}}\right),\; t'', \nu + \nu'\right).$$

$\underline{\mathsf{Mult}(\mathfrak{c}, \mathfrak{c}')}$: Given two ciphertexts representing messages $\mathbf{m}, \mathbf{m}' \in \mathbb{A}_2$, this algorithm forms a ciphertext encrypts the message $\mathbf{m} \cdot \mathbf{m}'$.

We begin by ensuring that the noise magnitude in both ciphertexts is smaller than the pre-set constant B (which is our base-line bound and is determined inEquation (12) below), performing modulus-switching as needed to ensure this condition. Then we bring both ciphertexts to the same level by reducing modulo the smaller of the two moduli (if needed). Once both ciphertexts have small noise magnitude and the same level we form the extended ciphertext (essentially performing the tensor product of the two) and apply key-switching to get back a normal ciphertext. A pseudo-code description of this procedure is given below.

$\underline{\mathsf{Mult}(\mathfrak{c}, \mathfrak{c}')}$:

1. While $\nu(\mathfrak{c}) > B$ do $\mathfrak{c} \leftarrow \mathsf{SwitchModulus}(\mathfrak{c})$; // $\nu(\mathfrak{c})$ is the noise estimate in $\mathfrak{c}$
2. While $\nu(\mathfrak{c}') > B$ do $\mathfrak{c}' \leftarrow \mathsf{SwitchModulus}(\mathfrak{c}')$; // $\nu(\mathfrak{c}')$ is the noise estimate in $\mathfrak{c}'$
3. Bring $\mathfrak{c}, \mathfrak{c}'$ to the same level t by reducing modulo the smaller of the two moduli
   Denote after modular reduction $\mathfrak{c} = ((c_0, c_1), t, \nu)$ and $\mathfrak{c}' = ((c'_0, c'_1), t, \nu')$
4. Set $(d_0, d_1, d_2) \leftarrow (c_0 \cdot c'_0\ ,\quad c_1 \cdot c'_0 + c_0 \cdot c'_1\ ,\quad -c_1 \cdot c'_1)$;
   Denote $\mathfrak{c}'' = ((d_0, d_1, d_2), t, \nu \cdot \nu')$
5. Output $\mathsf{SwitchKey}_{W[\mathfrak{s}^2 \to \mathfrak{s}]}(\mathfrak{c}'')$ // Convert to "normal" ciphertext

We stress that *the only place* where we force modulus switching is before the multiplication operation. In all other operations we allow the noise to grow, and it will be reduced back the first time it is input to a multiplication operation. We also note that we may need to apply modulus switching more than once before the noise is small enough.

$\underline{\mathsf{Scalar\text{-}Mult}(\mathfrak{c}, \alpha)}$: Given a ciphertext $\mathfrak{c} = (c_0, c_1, t, \nu)$ representing the message $\mathbf{m}$, and an element $\alpha \in \mathbb{A}_2$ (represented as a polynomial modulo 2 with coefficients in $\{-1, 0, 1\}$), this algorithm forms a ciphertext $\mathfrak{c}_m = (a_0, a_1, t_m, \nu_m)$ which encrypts the message $\mathbf{m}_m = \alpha \cdot \mathbf{m}$. This procedure is needed in our implementation of homomorphic AES, and is of more general interest in general computation over finite fields.

The algorithm makes use of a procedure $\mathsf{Randomize}(\alpha)$ which takes α and replaces each non-zero coefficients with a coefficients chosen at random from $\{-1, 1\}$. To multiply by α, we set $\beta \leftarrow \mathsf{Randomize}(\alpha)$ and then just multiply both c_0 and c_1 by β. Using the same argument as we used in Appendix A.5 for the distribution $\mathcal{HWT}(h)$, here too we can bound the norm of β by $\|\beta\|_\infty^{\mathsf{can}} \leq 6\sqrt{\mathsf{Wt}(\alpha)}$ where $\mathsf{Wt}(\alpha)$ is the number of nonzero coefficients of α. Hence we multiply the noise estimate by $6\sqrt{\mathsf{Wt}(\alpha)}$, and output the resulting ciphertext $\mathfrak{c}_m = (c_0 \cdot \beta,\ c_1 \cdot \beta,\ t,\ \nu \cdot 6\sqrt{\mathsf{Wt}(\alpha)})$.

$\mathsf{Automorphism}(\mathfrak{c}, \kappa)$: In the main body we explained how permutations on the plaintext slots can be realized via using elements $\kappa \in \mathcal{G}\mathsf{al}$; we also require the application of such automorphism to implement the Frobenius maps in our AES implementation.

For each κ that we want to use, we need to include in the public key the "matrix" $W[\kappa(\mathfrak{s}) \to \mathfrak{s}]$. Then, given a ciphertext $\mathfrak{c} = (c_0, c_1, t, \nu)$ representing the message $\mathbf{m}$, the function $\mathsf{Automorphism}(\mathfrak{c}, \kappa)$ produces a ciphertext $\mathfrak{c}' = (c_0', c_1', t, \nu')$ which represents the message $\kappa(\mathbf{m})$. We first set an "extended ciphertext" by setting

$$d_0 = \kappa(c_0), \quad d_1 \leftarrow 0, \quad \text{and } d_2 \leftarrow \kappa(c_1)$$

and then apply key switching to the extended ciphertext $((d_0, d_1, d_2), t, \nu)$ using the "matrix" $W[\kappa(\mathfrak{s}) \to \mathfrak{s}]$.

C Security Analysis and Parameter Settings

Below we derive the concrete parameters for use in our implementation. We begin in Appendix C.1 by deriving a lower-bound on the dimension N of the LWE problem underlying our key-switching matrices, as a function of the modulus and the noise variance. (This will serve as a lower-bound on $\phi(m)$ for our choice of the ring polynomial $\Phi_m(X)$.) Then in Appendix C.2 we derive a lower bound on the size of the largest modulus Q in our implementation, in terms of the noise variance and the dimension N. Then in Appendix C.3 we choose a value for the noise variance (as small as possible subject to some nominal security concerns), solve the somewhat circular constraints on N and Q, and set all the other parameters.

C.1 Lower-Bounding the Dimension

Below we apply to the LWE-security analysis of Lindner and Peikert [17], together with a few (arguably justifiable) assumptions, to analyze the dimension needed for different security levels. The analysis below assumes that we are given the modulus Q and noise variance σ^2 for the LWE problem (i.e., the noise is chosen from a discrete Gaussian distribution modulo Q with variance σ^2 in each coordinate). The goal is to derive a lower-bound on the dimension N required to get any given security level. The first assumption that we make, of course, is that the Lindner-Peikert analysis — which was done in the context of standard LWE — applies also for our ring-LWE case. We also make the following extra assumptions:

- We assume that (once σ is not too tiny), the security depends on the ratio Q/σ and not on Q and σ separately. Nearly all the attacks and hardness results in the literature support this assumption, with the exception of the Arora-Ge attack [2] (that works whenever σ is very small, regardless of Q).

- The analysis in [17] devised an experimental formula for the time that it takes to get a particular quality of reduced basis (i.e., the parameter δ of Gama and Nguyen [10]), then provided another formula for the advantage that the attack can derive from a reduced basis at a given quality, and finally used a computer program to solve these formulas for some given values of N and δ. This provides some

time/advantage tradeoff, since obtaining a smaller value of δ (i.e., higher-quality basis) takes longer time and provides better advantage for the attacker.

For our purposes we made the assumption that the best runtime/advantage ratio is achieved in the high-advantage regime. Namely we should spend basically all the attack running time doing lattice reduction, in order to get a good enough basis that will break security with advantage (say) 1/2. This assumption is consistent with the results that are reported in [17].

- Finally, we assume that to get advantage of close to 1/2 for an LWE instance with modulus Q and noise σ, we need to be able to reduce the basis well enough until the shortest vector is of size roughly Q/σ. Again, this is consistent with the results that are reported in [17].

Given these assumptions and the formulas from [17], we can now solve the dimension/security tradeoff analytically. Because of the first assumption we might as well simplify the equations and derive our lower bound on N for the case $\sigma = 1$, where the ratio Q/σ is equal to Q. (In reality we will use $\sigma \approx 4$ and increase the modulus by the same 2 bits).

Following Gama-Nguyen [10], recall that a reduced basis $B = (b_1|b_2|\dots|b_m)$ for a dimension-M, determinant-D lattice (with $\|b_1\| \le \|b_2\| \le \cdots \|b_M\|$), has quality parameter δ if the shortest vector in that basis has norm $\|b_1\| = \delta^M \cdot D^{1/M}$. In other words, the quality of B is defined as $\delta = \|b_1\|^{1/M}/D^{1/M^2}$. The time (in seconds) that it takes to compute a reduced basis of quality δ for a random LWE instance was estimated in [17] to be at least

$$\log(\text{time}) \;\ge\; 1.8/\log(\delta) - 110. \tag{7}$$

For a random Q-ary lattice of rank N, the determinant is exactly Q^N whp, and therefore a quality-δ basis has $\|b_1\| = \delta^M \cdot Q^{N/M}$. By our second assumption, we should reduce the basis enough so that $\|b_1\| = Q$, so we need $Q = \delta^M \cdot Q^{N/M}$. The LWE attacker gets to choose the dimension M, and the best choice for this attack is obtained when the right-hand-side of the last equality is minimized, namely for $M = \sqrt{N \log Q/\log\delta}$. This yields the condition

$$\log Q = \log(\delta^M Q^{N/M}) = M\log\delta + (N/M)\log Q = 2\sqrt{N\log Q\log\delta},$$

which we can solve for N to get $N = \log Q/4\log\delta$. Finally, we can use Equation (7) to express $\log\delta$ as a function of $\log(\text{time})$, thus getting $N = \log Q \cdot (\log(\text{time}) + 110)/7.2$. Recalling that in our case we used $\sigma = 1$ (so $Q/\sigma = Q$), we get our lower-bound on N in terms of Q/σ. Namely, to ensure a time/advantage ratio of at least 2^k, we need to set the rank N to be at least

$$N \;\ge\; \frac{\log(Q/\sigma)(k+110)}{7.2} \tag{8}$$

For example, the above formula says that to get 80-bit security level we need to set $N \ge \log(Q/\sigma)\cdot 26.4$, for 100-bit security level we need $N \ge \log(Q/\sigma)\cdot 29.1$, and for 128-bit security level we need $N \ge \log(Q/\sigma)\cdot 33.1$. We comment that these values are indeed consistent with the values reported in [17].

C.1.1 LWE with Sparse Key

The analysis above applies to "generic" LWE instance, but in our case we use very sparse secret keys (with only $h = 64$ nonzero coefficients, all chosen as ± 1). This brings up the question of whether one can get better attacks against LWE instances with a very sparse secret (much smaller than even the noise). The only attack that we could find that takes advantage of this sparse key is by applying the reduction technique of Applebaum et al. [1] to switch the key with part of the error vector, thus getting a smaller LWE error.

In a sparse-secret LWE we are given a random N-by-M matrix A (modulo Q), and also an M-vector $\mathbf{y} = [\mathbf{s}A + \mathbf{e}]_Q$. Here the N-vector $\mathbf{s}$ is our very sparse secret, and $\mathbf{e}$ is the error M-vector (which is also short, but not sparse and not as short as $\mathbf{s}$).

Below let A_1 denotes the first N columns of A, A_2 the next N columns, then A_3, A_4, etc. Similarly $\mathbf{e}_1, \mathbf{e}_2, \ldots$ are the corresponding parts of the error vector and $\mathbf{y}_1, \mathbf{y}_2, \ldots$ the corresponding parts of $\mathbf{y}$. Assuming that A_1 is invertible (which happens with high probability), we can transform this into an LWE instance with respect to secret $\mathbf{e}_1$, as follows:

We have $\mathbf{y}_1 = \mathbf{s}A_1 + \mathbf{e}_1$, or alternatively $A_1^{-1}\mathbf{y}_1 = \mathbf{s} + A_1^{-1}\mathbf{e}_1$. Also, for $i > 1$ we have $\mathbf{y}_i = \mathbf{s}A_i + \mathbf{e}_i$, which together with the above gives $A_iA_1^{-1}\mathbf{y}_1 - \mathbf{y}_i = A_iA_1^{-1}\mathbf{e}_1 - \mathbf{e}_i$. Hence if we denote

$$B_1 \stackrel{\text{def}}{=} A_1^{-1}, \quad \text{and for } i > 1 \; B_i \stackrel{\text{def}}{=} A_iA1^{-1},$$
$$\text{and similarly } \mathbf{z}_1 = A_1^{-1}\mathbf{y}_1, \quad \text{and for } i > 1 \;\; \mathbf{z}_i \stackrel{\text{def}}{=} A_iA_1^{-1}\mathbf{y}_i,$$

and then set $B \stackrel{\text{def}}{=} (B_1^t|B_2^t|B_3^t|\ldots)$ and $\mathbf{z} \stackrel{\text{def}}{=} (\mathbf{z}_1|\mathbf{z}_2|\mathbf{z}_3|\ldots)$, and also $\mathbf{f} = (\mathbf{s}|\mathbf{e}_2|\mathbf{e}_3|\ldots)$ then we get the LWE instance

$$\mathbf{z} = \mathbf{e}_1^t B + \mathbf{f}$$

with secret $\mathbf{e}_1^t$. The thing that makes this LWE instance potentially easier than the original one is that the first part of the error vector f is our sparse/small vector s, so the transformed instance has smaller error than the original (which means that it is easier to solve).

Trying to quantify the effect of this attack, we note that the optimal M value in the attack from Appendix C.1 above is obtained at $M = 2N$, which means that the new error vector is $\mathbf{f} = (\mathbf{s}|\mathbf{e}_2)$, which has Euclidean norm smaller than $\mathbf{e} = (\mathbf{e}_1|\mathbf{e}_2)$ by roughly a factor of $\sqrt{2}$ (assuming that $\|\mathbf{s}\| \ll \|\mathbf{e}_1\| \approx \|\mathbf{e}_2\|$). Maybe some further improvement can be obtained by using a smaller value for M, where the shorter error may outweigh the "non optimal" value of M. However, we do not expect to get major improvement this way, so it seems that the very sparse secret should only add maybe one bit to the modulus/noise ratio.

C.2 The Modulus Size

In this section we assume that we are given the parameter $N = \phi(m)$ (for our polynomial ring modulo $\Phi_m(X)$). We also assume that we are given the noise variance σ^2, the number of levels in the modulus chain L, an additional "slackness parameter" ξ (whose purpose is explained below), and the number of nonzero coefficients in the secret key h. Our goal is to devise a lower bound on the size of the largest modulus Q used in the public key, so as to maintain the functionality of the scheme.

Controlling the Noise. Driving the analysis in this section is a bound on the noise magnitude right after modulus switching, which we denote below by B. We set our parameters so that starting from ciphertexts with noise magnitude B, we can perform one level of fan-in-two multiplications, then one level of fan-in-ξ additions, followed by key switching and modulus switching again, and get the noise magnitude back to the same B.

- Recall that in the "reduced canonical embedding norm", the noise magnitude is at most multiplied by modular multiplication and added by modular addition, hence after the multiplication and addition levels the noise magnitude grows from B to as much as ξB^2.
- As we've seen in Appendix B.3, performing key switching scales up the noise magnitude by a factor of P and adds another noise term of magnitude upto $B_{\mathsf{Ks}} \cdot q_t$ (before doing modulus switching to scale it

back down). Hence starting from noise magnitude ξB^2, the noise grows to magnitude $P\xi B^2 + B_{\mathsf{Ks}} \cdot q_t$ (relative to the modulus Pq_t).

Below we assume that after key-switching we do modulus switching directly to a smaller modulus.

- After key-switching we can switch to the next modulus q_{t-1} to decrease the noise back to our bound B. Following the analysis from Appendix B.2, switching moduli from Q_t to q_{t-1} decreases the noise magnitude by a factor of $q_{t-1}/Q_t = 1/(P \cdot p_t)$, and then add a noise term of magnitude B_{scale}.

  Starting from noise magnitude $P\xi B^2 + B_{\mathsf{Ks}} \cdot q_t$ before modulus switching, the noise magnitude after modulus switching is therefore bounded whp by

$$\frac{P \cdot \xi B^2 + B_{\mathsf{Ks}} \cdot q_t}{P \cdot p_t} + B_{\mathsf{scale}} \;=\; \frac{\xi B^2}{p_t} + \frac{B_{\mathsf{Ks}} \cdot q_{t-1}}{P} + B_{\mathsf{scale}}$$

Using the analysis above, our goal next is to set the parameters B, P and the p_t's (as functions of N, σ, L, ξ and h) so that in every level t we get $\frac{\xi B^2}{p_t} + \frac{B_{\mathsf{Ks}} \cdot q_{t-1}}{P} + B_{\mathsf{scale}} \leq B$. Namely we need to satisfy at every level t the quadratic inequality (in B)

$$\frac{\xi}{p_t} B^2 \;-\; B \;+\; \Bigg(\underbrace{\frac{B_{\mathsf{Ks}} \cdot q_{t-1}}{P} + B_{\mathsf{scale}}}_{\text{denote this by } R_{t-1}} \Bigg) \;\leq\; 0\,. \tag{9}$$

Observe that (assuming that all the primes p_t are roughly the same size), it suffices to satisfy this inequality for the largest modulus $t = L-2$, since R_{t-1} increases with larger t's. Noting that $R_{L-3} > B_{\mathsf{scale}}$, we want to get this term to be as close to B_{scale} as possible, which we can do by setting P large enough. Specifically, to make it as close as $R_{L-3} = (1 + 2^{-n})B_{\mathsf{scale}}$ it is sufficient to set

$$P \;\approx\; 2^n \frac{B_{\mathsf{Ks}} q_{L-3}}{B_{\mathsf{scale}}} \;\approx\; 2^n \frac{9\sigma N q_{L-3}}{77\sqrt{N}} \;\approx\; 2^{n-3} q_{L-3} \cdot \sigma\sqrt{N}, \tag{10}$$

Below we set (say) $n = 8$, which makes it close enough to use just $R_{L-3} \approx B_{\mathsf{scale}}$ for the derivation below.

Clearly to satisfy Inequality (9) we must have a positive discriminant, which means $1 - 4\frac{\xi}{p_{L-2}} R_{L-3} \geq 0$, or $p_{L-2} \geq 4\xi R_{L-3}$. Using the value $R_{L-3} \approx B_{\mathsf{scale}}$, this translates into setting

$$p_1 \approx p_2 \cdots \approx p_{L-2} \approx 4\xi \cdot B_{\mathsf{scale}} \approx 308\xi\sqrt{N} \tag{11}$$

Finally, with the discriminant positive and all the p_i's roughly the same size we can satisfy Inequality (9) by setting

$$B \approx \frac{1}{2\xi/p_{L-2}} = \frac{p_{L-2}}{2\xi} \;\approx\; 2B_{\mathsf{scale}} \approx 154\sqrt{N}. \tag{12}$$

The Smallest Modulus. After evaluating our L-level circuit, we arrive at the last modulus $q_0 = p_0$ with noise bounded by ξB^2. To be able to decrypt, we need this noise to be smaller than $q_0/2c_m$, where c_m is the ring constant for our polynomial ring modulo $\Phi_m(X)$. For our setting, that constant is always below 40, so a sufficient condition for being able to decrypt is to set

$$q_0 = p_0 \;\approx\; 80\xi B^2 \;\approx\; 2^{20.9}\xi N \tag{13}$$

The Encryption Modulus. Recall that freshly encrypted ciphertxt have noise B_{clean} (as defined in Equation (6)), which is larger than our baseline bound B from above. To reduce the noise magnitude after the first modulus switching down to B, we therefore set the ratio $p_{L-1} = q_{L-1}/q_{L-2}$ so that $B_{\mathsf{clean}}/p_{L-1} + B_{\mathsf{scale}} \leq B$. This means that we set

$$p_{L-1} \;=\; \frac{B_{\mathsf{clean}}}{B - B_{\mathsf{scale}}} \approx \frac{146N + 858\sqrt{N}}{77\sqrt{N}} \approx 1.9\sqrt{N} + 11 \tag{14}$$

The Largest Modulus. Having set all the parameters, we are now ready to calculate the resulting bound on the largest modulus, namely $Q_{L-2} = q_{L-2} \cdot P$. Using Equations (11), and (13), we get

$$q_t \;=\; p_0 \cdot \prod_{i=1}^{t} p_i \;\approx\; (2^{20.9}\xi N) \cdot \left(308\xi\sqrt{N}\right)^t \;=\; 2^{20.9} \cdot 308^t \cdot \xi^{t+1} \cdot N^{t/2+1}. \tag{15}$$

Now using Equation (10) we have

$$\begin{aligned} P \;&\approx\; 2^5 q_{L-3}\sigma\sqrt{N} \;\approx\; 2^{25.9} \cdot 308^{L-3} \cdot \xi^{L-2} \cdot N^{(L-3)/2+1} \cdot \sigma\sqrt{N} \\ &\approx\; 2 \cdot 308^L \cdot \xi^{L-2}\sigma N^{L/2} \end{aligned}$$

and finally

$$\begin{aligned} Q_{L-2} \;=\; P \cdot q_{L-2} \;&\approx\; (2 \cdot 308^L \cdot \xi^{L-2}\sigma N^{L/2}) \cdot (2^{20.9} \cdot 308^{L-2} \cdot \xi^{L-1} \cdot N^{L/2}) \\ &\approx\; \sigma \cdot 2^{16.5L+5.4} \cdot \xi^{2L-3} \cdot N^L \end{aligned} \tag{16}$$

C.3 Putting It Together

We now have in Equation (8) a lower bound on N in terms of Q, σ and the security level k, and in Equation (16) a lower bound on Q with respect to N, σ and several other parameters. We note that σ is a free parameter, since it drops out when substituting Equation (16) in Equation (8). In our implementation we used $\sigma = 3.2$, which is the smallest value consistent with the analysis in [18].

For the other parameters, we set $\xi = 8$ (to get a small "wiggle room" without increasing the parameters much), and set the number of nonzero coefficients in the secret key at $h = 64$ (which is already included in the formulas from above, and should easily defeat exhaustive-search/birthday type of attacks). Substituting these values into the equations above we get

$$\begin{aligned} &p_0 \approx 2^{23.9}N, \quad p_i \approx 2^{11.3}\sqrt{N} \;\text{ for } i = 1, \ldots, L-2 \\ &P \approx 2^{11.3L-5}N^{L/2}, \;\text{ and } \quad Q_{L-2} \approx 2^{22.5L-3.6}\sigma N^L. \end{aligned}$$

Substituting the last value of Q_{L-2} into Equation (8) yields

$$N > \frac{(L(\log N + 23) - 8.5)(k + 110)}{7.2} \tag{17}$$

Targeting $k = 80$-bits of security and solving for several different depth parameters L, we get the results in the table below, which also lists approximate sizes for the primes p_i and P.

L	N	$\log_2(p_0)$	$\log_2(p_i)$	$\log_2(p_{L-1})$	$\log_2(P)$
10	9326	37.1	17.9	7.5	177.3
20	19434	38.1	18.4	8.1	368.8
30	29749	38.7	18.7	8.4	564.2
40	40199	39.2	18.9	8.6	762.2
50	50748	39.5	19.1	8.7	962.1
60	61376	39.8	19.2	8.9	1163.5
70	72071	40.0	19.3	9.0	1366.1
80	82823	40.2	19.4	9.1	1569.8
90	93623	40.4	19.5	9.2	1774.5

Choosing Concrete Values. Having obtained lower-bounds on $N = \phi(m)$ and other parameters, we now need to fix precise cyclotomic fields $\mathbb{Q}(\zeta_m)$ to support the algebraic operations we need. We have two situations we will be interested in for our experiments. The first corresponds to performing arithmetic on bytes in $\mathbb{F}_{2^8}$ (i.e. $n = 8$), whereas the latter corresponds to arithmetic on bits in $\mathbb{F}_2$ (i.e. $n = 1$). We therefore need to find an odd value of m, with $\phi(m) \approx N$ and m dividing $2^d - 1$, where we require that d is divisible by n. Values of m with a small number of prime factors are preferred as they give rise to smaller values of c_m. We also look for parameters which maximize the number of slots ℓ we can deal with in one go, and values for which $\phi(m)$ is close to the approximate value for N estimated above. When $n = 1$ we always select a set of parameters for which the ℓ value is at least as large as that obtained when $n = 8$.

	$n = 8$				$n = 1$			
L	m	$N = \phi(m)$	(d, ℓ)	c_K	m	$N = \phi(m)$	(d, ℓ)	c_K
10	11441	10752	(48,224)	3.60	11023	10800	(45,240)	5.13
20	34323	21504	(48,448)	6.93	34323	21504	(48,448)	6.93
30	31609	31104	(72,432)	5.15	32377	32376	(57,568)	1.27
40	54485	40960	(64,640)	12.40	42799	42336	(21,2016)	5.95
50	59527	51840	(72,720)	21.12	54161	52800	(60,880)	4.59
60	68561	62208	(72,864)	36.34	85865	63360	(60,1056)	12.61
70	82603	75264	(56,1344)	36.48	82603	75264	(56,1344)	36.48
80	92837	84672	(56,1512)	38.52	101437	85672	(42,2016)	19.13
90	124645	98304	(48,2048)	21.07	95281	94500	(45,2100)	6.22

D Further AES Implementation Methods

In this section we present our two other variants for implementing AES.

Byte Sliced Representation: In this representation we use sixteen distinct ciphertexts to represent a single state matrix, but within these sixteen ciphertexts we can also operate on s different state matrices. Again the underlying plaintext is assumed to consist of s slots, each containing a copy of the finite field $\mathbb{F}_{2^8}$. In this representation though there is no interaction between the slots, thus we operate with pure s-fold SIMD operations. As there is no interaction required between the slots we also therefore no longer require the use of our permutation networks. The SubBytes operation is implemented as above, but all the other ciphertext operations become essentially trivial, although the scalar multiplication in MixColumns will require us to consume another level. In pseudo code this becomes:

Input: ciphertexts $\mathfrak{c}_0, \ldots, \mathfrak{c}_{15}$ in level t — Level

// Compute SubBytes

For $i = 0, \ldots, 15$

// Compute $\mathfrak{c}_{254} = \mathfrak{c}_i^{-1}$

1. $\mathfrak{c}'_k \leftarrow \eta_k(\mathfrak{c}_i)$ for $k = 2, 4, 8, \ldots, 128$ — t
2. $\mathfrak{c}'_{254} \leftarrow ((\mathfrak{c}'_{128} \cdot \mathfrak{c}'_{64}) \cdot (\mathfrak{c}'_{32} \cdot \mathfrak{c}'_{16})) \cdot ((\mathfrak{c}'_8 \cdot \mathfrak{c}'_4) \cdot \mathfrak{c}'_2)$ // Three multiplication levels — $t - 3.0$

// Affine transformation over $\mathbb{F}_2$

3. $\mathfrak{c}'_k \leftarrow \eta_k(\mathfrak{c}'_{254})$ for $k = 1, 2, 4, 8, \ldots, 128$ — $t - 3.0$
4. $\mathfrak{c}^*_i \leftarrow \delta + \sum_{j=0}^{7} \gamma_j \cdot \mathfrak{c}'_{2^j}$ // Linear combination over $\mathbb{F}_{2^8}$ — $t - 3.5$

// Compute ShiftRows/MixColumns

For $i = 0, \ldots, 15$

1. $c \leftarrow i \bmod 4$, $r \leftarrow (i - c)/4$.
2. $a \leftarrow 4 \cdot r + (c + r \bmod 4)$. $b \leftarrow 4 \cdot (r + 1 \bmod 4) + (c + r + 1 \bmod 4)$.
3. $c \leftarrow 4 \cdot (r + 2 \bmod 4) + (c + r + 2 \bmod 4)$. $d \leftarrow 4 \cdot (r + 3 \bmod 4) + (c + r + 3 \bmod 4)$.
4. $\mathfrak{c}_i \leftarrow \mathfrak{c}^*_b + \mathfrak{c}^*_c + \mathfrak{c}^*_d + X \cdot (\mathfrak{c}^*_a + \mathfrak{c}^*_b)$ // Linear combination over $\mathbb{F}_{2^8}$ — $t - 4.0$.

Thus the expected minimum number of levels per round is four, in practice (again over many parameter sets) we find the average number of levels consumed per round is between 4.5 and 5.0. Note, we no longer need such a complicated set up operation. Namely we do not need to compute the key switching data for the $\sigma_{i,j}$ maps; we do however still need it for the Frobenius operations.

Bit Sliced Representation: For the bit sliced representation we represent the entire round function as a binary circuit, and we use 128 distinct ciphertexts (one per bit of the state matrix). However each set of 128 ciphertexts is able to represent a total of s distinct blocks. The main issue here is how to create a circuit for the round function which is as shallow, in terms of number of multiplication gates, as possible. Again the main issue is the SubBytes operation as all operations are essentially linear. To implement the SubBytes we used the "depth-16" circuit of Boyar and Peralta [3], which consumes four levels. The rest of the round function can be represented as a set of bit-additions, Thus, implementing this method means that we consumes a minimum of four levels on computing an entire round function. However, the extensive additions within the Boyar–Peralta circuit mean that we actually end up consuming a lot more. On average this translates into actually consuming between 5.0 and 10.0 levels per round.

E Scale(c, q_t, q_{t-1}) in dble-CRT Representation

Let $q_i = \prod_{j=0}^{i} p_j$, where the p_j's are primes that split completely in our cyclotomic field $\mathbb{A}$. We are given a $c \in \mathbb{A}_{q_t}$ represented via double-CRT – that is, it is represented as a "matrix" of its evaluations at the primitive m-th roots of unity modulo the primes $p_0, \ldots, p_t$. We want to modulus switch to q_{t-1} – i.e., scale down by a factor of p_t. Let's recall what this means: we want to output $c' \in \mathbb{A}$, represented via double-CRT format (as its matrix of evaluations modulo the primes $p_0, \ldots, p_{t-1}$), such that

1. $c' = c \bmod 2$.
2. c' is very close (in terms of its coefficient vector) to c/p_t.

In the main body we explained how this could be performed in dble-CRT representation. This made explicit use of the fact that the two ciphertexts need to be equivalent modulo two. If we wished to replace two with

a general prime p, then things are a bit more complicated. For completeness, although it is not required in our scheme, we present a methodology below. In this case, the conditions on $c^\dagger$ are as follows:

1. $c^\dagger = c \cdot p_t \bmod p$.
2. $c^\dagger$ is very close to c.
3. $c^\dagger$ is divisible by p_t.

As before, we set $c' \leftarrow c^\dagger / p_t$. (Note that for $p = 2$, we trivially have $c \cdot p_t = c \bmod p$, since p_t will be odd.)

This causes some complications, because we set $c^\dagger \leftarrow c + \delta$, where $\delta = -\bar{c} \bmod p_t$ (as before) but now $\delta = (p_t - 1) \cdot c \bmod p$. To compute such a δ, we need to know $c \bmod p$. Unfortunately, we don't have $c \bmod p$. One not-very-satisfying way of dealing with this problem is the following. Set $\hat{c} \leftarrow [p_t]_p \cdot c \bmod q_t$. Now, if c encrypted m, then $\hat{c}$ encrypts $[p_t]_p \cdot m$, and $\hat{c}$'s noise is $[p_t]_p < p/2$ times as large. It is obviously easy to compute $\hat{c}$'s double-CRT format from c's. Now, we set $c^\dagger$ so that the following is true:

1. $c^\dagger = \hat{c} \bmod p$.
2. $c^\dagger$ is very close to $\hat{c}$.
3. $c^\dagger$ is divisible by p_t.

This is easy to do. The algorithm to output $c^\dagger$ in double-CRT format is as follows:

1. Set $\bar{c}$ to be the coefficient representation of $\hat{c} \bmod p_t$. (Computing this requires a single "small FFT" modulo the prime p_t.)
2. Set δ to be the polynomial with coefficients in $(-p_t, p_t]$ such that $\delta = 0 \bmod p$ and $\delta = -\bar{c} \bmod p_t$.
3. Set $c^\dagger = \hat{c} + \delta$, and output $c^\dagger$'s double-CRT representation.
   (a) We already have $\hat{c}$'s double-CRT representation.
   (b) Computing δ's double-CRT representation requires t "small FFTs" modulo the p_j's.