CRYSTALS-Kyber

Algorithm Specifications And Supporting Documentation

(version 3.02)

Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint,
Vadim Lyubashevsky, John M. Schanck, Peter Schwabe, Gregor Seiler, Damien Stehlé

August 4, 2021

Changelog

In the following we list the changes from Kyber as submitted to round-2 of the the NIST PQC project in April 2019 and Kyber with round-3 tweaks together with brief explanations of the motivation for the changes.

Changes to the core Kyber design

- Increase noise parameter for Kyber512

  In the round-2 submission of Kyber, the decryption error for Kyber512 was rather conservative, while at the same time, there were requests to increase the Core-SVP hardness of this parameter set. It therefore made sense to increase the binomial error distribution for Kyber512. Instead of increasing it from 2 to 3 across the board, we chose to use the fact that the rounding of the ciphertext coefficients to 10-bit integers also effectively adds some noise. We therefore increased the binomial noise from 2 to 3 of the parameters in the key generation procedure and of the parameter $\mathbf{r}$ in the encryption procedure, but not of errors $\mathbf{e}_1$ and e_2.

  Relying on the rounding noise to add error is akin to the LWR assumption, but our reliance on it is quite small. First, it only adds 6 bits of Core-SVP hardness, and second, we are adding noise and rounding, which presumably has less algebraic structure than just rounding. In short, without the LWR assumption, our new parameter set for Kyber512 still has 112 bits of core-SVP hardness as before, while with a weak version of the LWR assumption, it has 118 bits.

- Reduce ciphertext compression of Kyber512

  By making the noise larger, the probability of decryption failures also increases. To compensate, we drop one less bit in the "second" ciphertext element, which increases our ciphertext size from 736 bytes to 768 bytes, for a 2^{-139} decryption error.

- More efficient uniform sampling of the public matrix $\mathbf{A}$

  Instead of sampling uniformly-random integers modulo 3329 by using rejection sampling on a 2-byte integer, we now use rejection sampling on a 12-bit integer. While the new rejection rate per coefficient is higher (i.e. $\approx 20\%$), the total number of required bits and the running time of key generation are noticeably smaller.

Changes to the specification and supporting documentation

- Updated specification to match the round-3 parameters.
- Updated performance numbers.
- ARM Cortex-M4 numbers are now included in the performance-analysis section. The corresponding software (to be dropped into pqm4 [59]) is included in the submission package.
- A major update of the security analysis, in particular a much more detailed analysis of the stateof-the art understanding of attacks beyond "core-SVP" hardness and a updated discussion of attacks exploiting decryption failures.

Contents

1 Written specification

Kyber is an IND-CCA2-secure key-encapsulation mechanism (KEM), which has first been described in [24]. The security of Kyber is based on the hardness of solving the learning-with-errors problem in module lattices (MLWE problem [67]). The construction of Kyber follows a two-stage approach: we first introduce an INDCPA-secure public-key encryption scheme encrypting messages of a fixed length of 32 bytes, which we call Kyber.CPAPKE. We then use a slightly tweaked Fujisaki–Okamoto (FO) transform [47] to construct the IND-CCA2-secure KEM. Whenever we want to emphasize that we are speaking about the IND-CCA2-secure KEM, we will refer to it as Kyber.CCAKEM.

In Subsection 1.1 we give preliminaries and fix notation. In Subsection 1.2 we give a full specification of Kyber.CPAPKE. Subsection 1.3 gives details of the transform that we use in Kyber to obtain Kyber.CCAKEM from Kyber.CPAPKE. Subsection 1.4 lists the parameters that we propose for different security levels. Finally, Subsection 1.5 explains the design rationale behind Kyber.

1.1 Preliminaries and notation.

Bytes and byte arrays. Inputs and outputs to all API functions of Kyber are byte arrays. To simplify notation, we denote by by B the set $\{0,\ldots,255\}$, i.e., the set of 8-bit unsigned integers (bytes). Consequently we denote by B^k the set of byte arrays of length k and by B^* the set of byte arrays of arbitrary length (or byte streams). For two byte arrays a and b we denote by $(a\mathrm{k}b)$ the concatenation of a and b. For a byte array a we denote by $a + k$ the byte array starting at byte k of a (with indexing starting at zero). For example, let a by a byte array of length $\grave{}$, let b be another byte array and let $c = (a\mathrm{k}b)$ be the concatenation of a and b; then $b = a + \grave{}$. When it is more convenient to work with an array of bits than an array of bytes we make this conversion explicit via the BytesToBits function that takes as input an array of $\grave{}$ bytes and produces as output an array of $8\grave{}$ bits. Bit β_i at position i of the output bit array is obtained from byte $b_{i/8}$ at position $i/8$ of the input array by computing $\beta_i = \left((b_{i/8}/2^{(i \mod 8)}) \mod 2\right)$.

Polynomial rings and vectors. We denote by R the ring $\mathrm{Z}[X]/(X^n+1)$ and by R_q the ring $\mathrm{Z}_q[X]/(X^n+1)$, where $n = 2^{n_0-1}$ such that $X^n + 1$ is the 2^{n_0}-th cyclotomic polynomial. Throughout this document, the values of n, n^0 and q are fixed to $n = 256$, $n^0 = 9$, and $q = 3329$. Regular font letters denote elements in R or R_q (which includes elements in Z and Z_q) and bold lower-case letters represent vectors with coefficients in R or R_q. By default, all vectors will be column vectors.

Bold upper-case letters are matrices. For a vector $\mathbf{v}$ (or matrix $\mathbf{A}$), we denote by $\mathbf{v}^T$ (or $\mathbf{A}^T$) its transpose. For a vector $\mathbf{v}$ we write $\mathbf{v}[i]$ to denote it's i-th entry (with indexing starting at zero); for a matrix $\mathbf{A}$ we write $\mathbf{A}[i][j]$ to denote the entry in row i, column j (again, with indexing starting at zero).

Modular reductions. For an even (resp. odd) positive integer α, we define $r' = r \bmod^{\pm} \alpha$ to be the unique element r' in the range $-\frac{\alpha}{2} < r' \leq \frac{\alpha}{2}$ (resp. $-\frac{\alpha-1}{2} \leq r' \leq \frac{\alpha-1}{2}$) such that $r' = r \bmod \alpha$. For any positive integer α, we define $r' = r \bmod^{+} \alpha$ to be the unique element r' in the range $0 \leq r' < \alpha$ such that $r' = r \bmod \alpha$. When the exact representation is not important, we simply write $r \bmod \alpha$.

Rounding. For an element $x \in \mathbb{Q}$ we denote by $\lfloor x \rceil$ rounding of x to the closest integer with ties being rounded up.

Sizes of elements. For an element $w \in \mathbb{Z}_q$, we write $\|w\|_\infty$ to mean $|w \bmod^{\pm} q|$. We now define the ℓ_∞ and ℓ_2 norms for $w = w_0 + w_1 X + \ldots + w_{n-1} X^{n-1} \in R$:

$$\|w\|_\infty = \max_i \|w_i\|_\infty, \quad \|w\| = \sqrt{\|w_0\|_\infty^2 + \ldots + \|w_{n-1}\|_\infty^2}.$$

Similarly, for $\mathbf{w} = (w_1, \ldots, w_k) \in R^k$, we define

$$\|\mathbf{w}\|_\infty = \max_i \|w_i\|_\infty, \quad \|\mathbf{w}\| = \sqrt{\|w_1\|^2 + \ldots + \|w_k\|^2}.$$

Sets and Distributions. For a set S, we write $s \leftarrow S$ to denote that s is chosen uniformly at random from S. If S is a probability distribution, then this denotes that s is chosen according to the distribution S.

Compression and Decompression. We now define a function $\text{Compress}_q(x,d)$ that takes an element $x \in \mathbb{Z}_q$ and outputs an integer in $\{0,\ldots,2^d - 1\}$, where $d < \lfloor \log_2(q) \rfloor$. We furthermore define a function Decompress_q, such that

$$x_0 = \text{Decompress}_q\left(\text{Compress}_q(x,d), d\right) \tag{1}$$

is an element close to x – more specifically

$$|x' - x \bmod^{\pm} q| \leq B_q := \left\lceil \frac{q}{2^{d+1}} \right\rfloor. \tag{2}$$

The functions satisfying these requirements are defined as:

$$\text{Compress}_q(x,d) = \lfloor (2^d/q) \cdot x \rceil \bmod^{+} 2^d, \quad \text{Decompress}_q(x,d) = \lfloor (q/2^d) \cdot x \rceil.$$

When Compress_q or Decompress_q is used with $x \in R_q$ or $\mathbf{x} \in R_q^k$, the procedure is applied to each coefficient individually.

The main reason for defining the Compress_q and Decompress_q functions is to be able to discard some low-order bits in the ciphertext which do not have much effect on the correctness probability of decryption – thus reducing the size of ciphertexts.

The Compress_q and Decompress_q are also used for a purpose other than compression – namely to perform the usual LWE error correction during encryption and decryption. More precisely, in line 20 of the encryption procedure (Algorithm 5) the Decompress_q function is used to create error tolerance gaps by sending the message bit 0 to 0 and 1 to $\lfloor q/2 \rceil$. Later, on line 4 of the decryption procedure (Algorithm 6), the Compress_q function is used to decrypt to a 1 if $v - \mathbf{s}^T\mathbf{u}$ is closer to $\lfloor q/2 \rceil$ than to 0, and decrypt to a 0 otherwise.

Symmetric primitives. The design of Kyber makes use of a pseudorandom function PRF: $\mathrm{B}^{32} \times \mathrm{B} \to \mathrm{B}^*$ and of an extendable output function XOF: $\mathrm{B}^* \times \mathrm{B} \times \mathrm{B} \to \mathrm{B}^*$. Kyber also makes use of two hash functions H: $\mathrm{B}^* \to \mathrm{B}^{32}$ and G: $\mathrm{B}^* \to \mathrm{B}^{32} \times \mathrm{B}^{32}$ and of a key-derivation function KDF: $\mathrm{B}^* \to \mathrm{B}^*$.

NTTs, multiplication, and bitreversed order. A very efficient way to perform multiplications in R_q is via the so-called *number-theoretic transform* (NTT).

For our prime $q = 3329$ with $q-1 = 2^8 \cdot 13$, the base field Z_q contains primitive 256-th roots of unity but not primitive 512-th roots. Therefore, the defining polynomial $X^{256} + 1$ of R factors into 128 polynomials of degree 2 modulo q and the NTT of a polynomial $f \in R_q$ is a vector of 128 polynomials of degree one. Simple in-place implementations of the NTT without reordering outputs these polynomials in bit-reversed order and we define the NTT in this way. Concretely, let $\zeta = 17$ be the first primitive 256-th root of unity modulo q, and $\{\zeta,\zeta^3,\zeta^5,\ldots,\zeta^{255}\}$ the set of all the 256-th roots of unity. The polynomial $X^{256} + 1$ can therefore be written as

$$X^{256} + 1 = \prod_{i=0}^{127}(X^2 - \zeta^{2i+1}) = \prod_{i=0}^{127}(X^2 - \zeta^{2\mathrm{br}_7(i)+1}),$$

where $\mathrm{br}_7(i)$ for $i = 0,\ldots,127$ is the bit reversal of the unsigned 7-bit integer i. This latter ordering of the factors is useful for compatibility with the idiosyncrasies of AVX instructions. Then the NTT of $f \in R_q$ is given by

$$(f \bmod X^2 - \zeta^{2\mathrm{br}_7(0)+1},\ldots,f \bmod X^2 - \zeta^{2\mathrm{br}_7(127)+1}). \tag{3}$$

This vector of linear polynomials is then serialized to a vector in Z_q^{256} in the canonical way. Moreover, in order to not introduce additional data types and facilitate in-place implementations of the NTT we define NTT: $R_q \to R_q$ to be the bijection that maps $f \in R_q$ to the polynomial with the aforementioned coefficient vector. Hence,

$$\mathrm{NTT}(f) = \hat{f} = \hat{f}_0 + \hat{f}_1 X + \cdots + \hat{f}_{255} X^{255}$$

with

$$\hat{f}_{2i} = \sum_{j=0}^{127} f_{2j}\zeta^{(2\mathrm{br}_7(i)+1)j}, \tag{4}$$

$$\hat{f}_{2i+1} = \sum_{j=0}^{127} f_{2j+1}\zeta^{(2\mathrm{br}_7(i)+1)j}. \tag{5}$$

We would like to stress that even though we write $\hat{f}$ as a polynomial in R_q, it has no algebraic meaning as such. The natural algebraic representation of $\mathrm{NTT}(f) = \hat{f}$ is as 128 polynomials of degree 1 as in (3) using the definitions for $\hat{f}_i$ from (4) and (5). That is,

$$\mathrm{NTT}(f) = \hat{f} = (\hat{f}_0 + \hat{f}_1 X, \hat{f}_2 + \hat{f}_3 X, \ldots, \hat{f}_{254} + \hat{f}_{255} X).$$

Using NTT and its inverse NTT^{-1} we can compute the product $f \cdot g$ of two elements $f,g \in R_q$ very efficiently as $\mathrm{NTT}^{-1}(\mathrm{NTT}(f) \circ \mathrm{NTT}(g))$ where $\mathrm{NTT}(f) \circ \mathrm{NTT}(g) = \hat{f} \circ \hat{g} = \hat{h}$ denotes the basecase multiplication consisting of the 128 products

$$\hat{h}_{2i} + \hat{h}_{2i+1} X = (\hat{f}_{2i} + \hat{f}_{2i+1} X)(\hat{g}_{2i} + \hat{g}_{2i+1} X) \bmod X^2 - \zeta^{2\mathrm{br}_7(i)+1}$$

of linear polynomials.

When we apply NTT or NTT^{-1} to a vector or matrix of elements of R_q, then this means that the respective operation is applied to each entry individually. When we apply $\circ$ to matrices or vectors it means that we perform a usual matrix multiplication, but that the individual products of entries are the above basecase multiplications.

Throughout the document we will write NTT and NTT^{-1} whenever we refer to the concrete functions as defined above and use normal-font NTT whenever we refer to the general technique.

Uniform sampling in R_q. Kyber uses a deterministic approach to sample elements in R_q that are statistically close to a uniformly random distribution. For this sampling we use a function Parse: $\text{B}^* \to R_q$, which receives as input a byte stream $B = b_0,b_1,b_2,\ldots$ and computes the NTT-representation $a\hat{} = a\hat{}_0 + a\hat{}_1X + \cdots + a\hat{}_{n-1}X^{n-1} \in R_q$ of $a \in R_q$. Parse is described in Algorithm 1 (note that this description assumes that $q = 3329$).

Algorithm 1 Parse: $\text{B}^* \to R_q{}^n$

Input: Byte stream $B = b_0,b_1,b_2 \cdots \in \text{B}^*$
Output: NTT-representation $a\hat{} \in R_q$ of $a \in R_q$ $i := 0$
$j := 0$ while $j < n$ do $d_1 := b_i + 256 \cdot (b_{i+1} \bmod^+ 16)$
$d_2 := \text{b}b_{i+1}/16\text{c} + 16 \cdot b_{i+2}$ if
$d_1 < q$ then $a\hat{}_j := d_1$
$j := j + 1$ end if
if $d_2 < q$ and $j < n$ then $a\hat{}_j := d_2$
$j := j + 1$
end if $i := i + 3$
end while
return $a\hat{}_0 + a\hat{}_1X + \cdots + a\hat{}_{n-1}X^{n-1}$

The intuition behind the function Parse is that if the input byte array is statistically close to a uniformly random byte array, then the output polynomial is statistically close to a uniformly random element of R_q. It represents a uniformly random polynomial in R_q because NTT is bijective and thus maps polynomials with uniformly random coefficients to polynomials with again uniformly random coefficients.

Sampling from a binomial distribution. Noise in Kyber is sampled from a centered binomial distribution B_η for $\eta = 2$ or $\eta = 3$. We define B_η as follows:

$$\text{Sample } (a^1,\ldots,a_\eta,b_1,\ldots,b_\eta) \leftarrow \{0,1\}^{2\eta}$$

$$\text{and output } \sum_{i=1}^{\eta}(a_i - b_i).$$

When we write that a polynomial $f \in R_q$ or a vector of such polynomials is sampled from B_η, we mean that each coefficient is sampled from B_η.

For the specification of Kyber we need to define how a polynomial $f \in R_q$ is sampled according to B_η deterministically from 64η bytes of output of a pseudorandom function (we fix $n = 256$ in this description). This is done by the function CBD (for "centered binomial distribution") defined as described in Algorithm 2.

1.2 Specification of Kyber.CPAPKE

Kyber.CPAPKE is similar to the LPR encryption scheme that was introduced (for Ring-LWE) by Lyubashevsky, Peikert, and Regev in the presentation of [72] at Eurocrypt 2010 [73]; the description is also in the full version of the paper [74, Sec. 1.1]. The roots of this scheme go back to the first LWE-based encryption

Algorithm 2 $\mathrm{CBD}_\eta : \mathrm{B}^{64\eta} \to R_q$

Input: Byte array $B = (b_0, b_1, \ldots, b_{64\eta-1}) \in \mathrm{B}^{64\eta}$

Output: Polynomial $f \in R_q$

$(\beta_0, \ldots, \beta_{512\eta-1}) :=$ BytesToBits(B) for i from 0 to 255 do

$a := \sum_{j=0}^{\eta-1} \beta_{2i\eta+j}$

$b := \sum_{j=0}^{\eta-1} \beta_{2i\eta+\eta+j}$ $f_i := a - b$

end for return $f_0 + f_1 X + f_2 X^2 + \cdots + f_{255} X^{255}$

Encoding and decoding. There are two data types that Kyber needs to serialize to byte arrays: byte arrays and (vectors of) polynomials. Byte arrays are trivially serialized via the identity, so we need to define how we serialize and deserialize polynomials. In Algorithm 3 we give a pseudocode description of the function Decode$_\ell$, which deserializes an array of 32ℓ bytes into a polynomial $f = f_0 + f_1 X + \cdots + f_{255} X^{255}$ (we again fix $n = 256$ in this description) with each coefficient f_i in $\{0, \ldots, 2^\ell - 1\}$. We define the function Encode$_\ell$ as the inverse of Decode$_\ell$. Whenever we apply Encode$_\ell$ to a vector of polynomials we encode each polynomial individually and concatenate the output byte arrays.

Algorithm 3 $\mathrm{Decode}_\ell : \mathrm{B}^{32\ell} \to R_q$

Input: Byte array $B \in \mathrm{B}^{32\ell}$

Output: Polynomial $f \in R_q$

$(\beta_0, \ldots, \beta_{256\ell-1}) :=$ BytesToBits(B) for i from 0 to 255 do

$f_i := \sum_{j=0}^{\ell} \beta_{i\ell+j} 2^j$ end for return $f_0 + f_1 X + f_2 X^2 + \cdots + f_{255} X^{255}$

scheme presented by Regev in [94, 95], with the main difference being that the underlying ring is not $\mathbb{Z}_q$ and both the secret and the error vectors have small coefficients. The idea of using a polynomial ring (instead of $\mathbb{Z}_q$) goes back to the NTRU cryptosystem presented by Hoffstein, Pipher, and Silverman in [54], while the symmetry between the secret and the error was already employed in very similar cryptographic schemes in [8, 71] with the security justification from [12].

The main difference from the LPR encryption scheme is to use Module-LWE instead of Ring-LWE. Also, we adopt the approach taken by Alkım, Ducas, Pöppelmann and Schwabe in [10] for the generation of the public matrix $\mathbf{A}$. Furthermore, we shorten ciphertexts by rounding off the low bits as in learning-withrounding-based schemes [15, Eq. 2.1], which is a common technique for reducing ciphertext size also in LWE-based schemes (c.f. [84, 90]).

Parameters. Kyber.CPAPKE is parameterized by integers n, k, q, η_1, η_2, d_u, and d_v. As stated before, throughout this document n is always 256 and q is always 3329.

Using the notation of Subsection 1.1 we give the definition of key generation, encryption, and decryption of the Kyber.CPAPKE public-key encryption scheme in Algorithms 4, 5, and 6. A more high-level view of these algorithms is given in the comments.

Algorithm 4 Kyber.CPAPKE.KeyGen(): key generation

Output: Secret key $sk \in \mathrm{B}^{12 \cdot k \cdot n/8}$

Output: Public key $pk \in \mathrm{B}^{12 \cdot k \cdot n/8+32}$

1: $d \leftarrow \mathrm{B}^{32}$
2: $(\rho,\sigma) := \mathrm{G}(d)$
3: $N := 0$
4: for i from 0 to $k-1$ do . Generate matrix $\hat{\mathbf{A}} \in R_q^{k \times k}$ in NTT domain
5: for j from 0 to $k-1$ do
6: $\hat{\mathbf{A}}[i][j] := \mathrm{Parse}(\mathrm{XOF}(\rho,j,i))$
7: end for
8: end for
9: for i from 0 to $k-1$ do . Sample $\mathbf{s} \in R_q^k$ from B_{η_1}
10: $\mathbf{s}[i] := \mathrm{CBD}_{\eta_1}(\mathrm{PRF}(\sigma,N))$
11: $N := N + 1$
12: end for
13: for i from 0 to $k-1$ do . Sample $\mathbf{e} \in R_q^k$ from B_{η_1}
14: $\mathbf{e}[i] := \mathrm{CBD}_{\eta_1}(\mathrm{PRF}(\sigma,N))$
15: $N := N + 1$
16: end for
17: $\hat{\mathbf{s}} := \mathrm{NTT}(\mathbf{s})$
18: $\hat{\mathbf{e}} := \mathrm{NTT}(\mathbf{e})$
19: $\hat{\mathbf{t}} := \hat{\mathbf{A}} \circ \hat{\mathbf{s}} + \hat{\mathbf{e}}$
20: $pk := (\mathrm{Encode}_{12}(\hat{\mathbf{t}} \bmod^+ q) \| \rho)$. $pk := \mathbf{As} + \mathbf{e}$
21: $sk := \mathrm{Encode}_{12}(\hat{\mathbf{s}} \bmod^+ q)$. $sk := \mathbf{s}$
22: return (pk,sk)

Algorithm 5 Kyber.CPAPKE.Enc(pk,m,r): encryption

Input: Public key $pk \in \mathrm{B}^{12 \cdot k \cdot n/8+32}$ Input:

Message $m \in \mathrm{B}^{32}$

Input: Random coins $r \in \mathrm{B}^{32}$

Output: Ciphertext $c \in \mathrm{B}^{d_u \cdot k \cdot n/8+d_v \cdot n/8}$

1: $N := 0$
2: $\hat{\mathbf{t}} := \mathrm{Decode}_{12}(pk)$
3: $\rho := pk + 12 \cdot k \cdot n/8$
4: for i from 0 to $k-1$ do . Generate matrix $\hat{\mathbf{A}} \in R_q^{k \times k}$ in NTT domain
5: for j from 0 to $k-1$ do
6: $\hat{\mathbf{A}}^T[i][j] := \mathrm{Parse}(\mathrm{XOF}(\rho,i,j))$
7: end for

8: end for
9: for i from 0 to $k-1$ do 10: $\mathbf{r}[i]$:= . Sample $\mathbf{r} \in R_q^k$ from B_{η_1}
CBD$_{\eta_1}$(PRF(r,N))
11: $N := N + 1$
12: end for
13: for i from 0 to $k-1$ do 14: $\mathbf{e}_1[i]$:= . Sample $\mathbf{e}_1 \in R_q^k$ from B_{η_2}
CBD$_{\eta_2}$(PRF(r,N))
15: $N := N + 1$
16: end for
17: e_2 := CBD$_{\eta_2}$(PRF(r,N)) . Sample $e_2 \in R_q$ from B_{η_2}
18: $\hat{\mathbf{r}}$:= NTT$(\mathbf{r})$
19: $\mathbf{u}$:= NTT$^{-1}(\hat{\mathbf{A}}^T \circ \hat{\mathbf{r}}) + \mathbf{e}_1$. $\mathbf{u} := \mathbf{A}^T\mathbf{r} + \mathbf{e}_1$
20: v := NTT$^{-1}(\hat{\mathbf{t}}^T \circ \hat{\mathbf{r}}) + e_2$ + Decompress$_q$(Decode$_1(m),1)$. $v := \mathbf{t}^T\mathbf{r} + e_2$ + Decompress$_q(m,1)$
21: c_1 := Encode$_{d_u}$(Compress$_q(\mathbf{u},d_u)$)
22: c_2 := Encode$_{d_v}$(Compress$_q(v,d_v)$)
23: return $c = (c_1 \| c_2)$. c := (Compress$_q(\mathbf{u},d_u)$,Compress$_q(v,d_v)$)

Algorithm 6 Kyber.CPAPKE.Dec(sk,c): decryption

Input: Secret key $sk \in \mathrm{B}^{12\cdot k\cdot n/8}$
Input: Ciphertext $c \in \mathrm{B}^{d_u\cdot k\cdot n/8+d_v\cdot n/8}$
Output: Message $m \in \mathrm{B}^{32}$
1: $\mathbf{u}$:= Decompress$_q$(Decode$_{d_u}(c),d_u$)
2: v := Decompress$_q$(Decode$_{d_v}(c + d_u \cdot k \cdot n/8),d_v$)
3: $\hat{\mathbf{s}}$:= Decode$_{12}(sk)$
4: m := Encode$_1$(Compress$_q(v$ − NTT$^{-1}(\hat{\mathbf{s}}^T \circ$ NTT$(\mathbf{u})),1))$. m := Compress$_q(v - \mathbf{s}^T\mathbf{u},1))$
5: return m

1.3 Specification of Kyber.CCAKEM

We construct the Kyber.CCAKEM IND-CCA2-secure KEM from the IND-CPA-secure public-key encryption scheme described in the previous subsection via a slightly tweaked Fujisaki–Okamoto transform [47]. In Algorithms 7, 8, and 9 we define key generation, encapsulation, and decapsulation of Kyber.CCAKEM.

Algorithm 7 Kyber.CCAKEM.KeyGen()

Output: Public key $pk \in \mathrm{B}^{12 \cdot k \cdot n/8+32}$

Output: Secret key $sk \in \mathrm{B}^{24 \cdot k \cdot n/8+96}$

1: $z \leftarrow \mathrm{B}^{32}$

2: $(pk, sk^0) := \text{Kyber.CPAPKE.KeyGen}()$

3: $sk := (sk^0 \| pk \| \mathrm{H}(pk) \| z)$

4: return (pk, sk)

Algorithm 8 Kyber.CCAKEM.Enc(pk)

Input: Public key $pk \in \mathrm{B}^{12 \cdot k \cdot n/8+32}$

Output: Ciphertext $c \in \mathrm{B}^{d_u \cdot k \cdot n/8+d_v \cdot n/8}$

Output: Shared key $K \in \mathrm{B}^*$

1: $m \leftarrow \mathrm{B}^{32}$

2: $m \leftarrow \mathrm{H}(m)$. Do not send output of system RNG

3: $(K, \bar{r}) := \mathrm{G}(m \| \mathrm{H}(pk))$

4: $c := \text{Kyber.CPAPKE.Enc}(pk, m, r)$

5: $K := \mathrm{KDF}(\bar{K} \| \mathrm{H}(c))$

6: return (c, K)

Algorithm 9 Kyber.CCAKEM.Dec(c, sk)

Input: Ciphertext $c \in \mathrm{B}^{d_u \cdot k \cdot n/8+d_v \cdot n/8}$ Input: Secret key $sk \in \mathrm{B}^{24 \cdot k \cdot n/8+96}$

Output: Shared key $K \in \mathrm{B}^*$

1: $pk := sk + 12 \cdot k \cdot n/8$

2: $h := sk + 24 \cdot k \cdot n/8 + 32 \in \mathrm{B}^{32}$

3: $z := sk + 24 \cdot k \cdot n/8 + 64$

4: $m^0 := \text{Kyber.CPAPKE.Dec}(sk, c)$

5: $(\bar{K}^0, r^0) := \mathrm{G}(m^0 \| h)$

6: $c^0 := \text{Kyber.CPAPKE.Enc}(pk, m^0, r^0)$

7: if $c = c^0$ then

8: return $K := \mathrm{KDF}(\bar{K}^0 \| \mathrm{H}(c))$

9: else

10: return $K := \mathrm{KDF}(z \| \mathrm{H}(c))$

11: end if

12: return K

1.4 Kyber parameter sets

We define three parameter sets for Kyber, which we call Kyber512, Kyber768, and Kyber1024. The parameters are listed in Table 1. Note that the table also lists the derived parameter δ, which is the probability that decapsulation of a valid Kyber.CCAKEM ciphertext fails. The parameters were obtained via the following approach:

Table 1: Parameter sets for Kyber

	n	k	q	η_1	η_2	(d_u,d_v)	δ
Kyber512	256	2	3329	3	2	(10,4)	2^{-139}
Kyber768	256	3	3329	2	2	(10,4)	2^{-164}
Kyber1024	256	4	3329	2	2	(11,5)	2^{-174}

- n is set to 256 because the goal is to encapsulate keys with 256 bits of entropy (i.e., use a plaintext size of 256 bits in Kyber.CPAPKE.Enc). Smaller values of n would require to encode multiple key bits into one polynomial coefficient, which requires lower noise levels and therefore lowers security. Larger values of n would reduce the capability to easily scale security via parameter k.
- We choose q as a small prime satisfying $n \mid (q-1)$; this is required to enable the fast NTT-based multiplication. There are two smaller primes for which this property holds, namely 257 and 769. However, for those primes we would not be able to achieve negligible failure probability required for CCA security, so we chose the next largest, i.e., $q = 3329$.
- k is selected to fix the lattice dimension as a multiple of n; changing k is the main mechanism in Kyber to scale security (and as a consequence, efficiency) to different levels.
- The remaining parameters η_1, η_2, d_u and d_v were chosen to balance between security (see Section 4), ciphertext size, and failure probability. Note that all three parameter sets achieve a failure probability of $< 2^{-128}$ with some margin. We discuss this in more detail in Subsections 1.5 and 5.5.
- The parameter η_1 defines the noise of $\mathbf{s}$ and $\mathbf{e}$ in Algorithm 4 and of $\mathbf{r}$ in Algorithm 5. The parameter η_2 defines the noise of $\mathbf{e}_1$ and e_2 in Algorithm 5. We discuss the reason for setting $\eta_1 > \eta_2$ in the Kyber512 parameter set in Section 1.5.

The failure probability δ is computed with the help of the Kyber.py Python script which is available online at https://github.com/pq-crystals/security-estimates. For the theoretical background of that script see [24, Theorem 1].

Instantiating PRF, XOF, H, G, and KDF. What is still missing to complete the specification of Kyber is the instantiation of the symmetric primitives. We instantiate all of those primitives with functions from the FIPS-202 standard [81] as follows:

- We instantiate XOF with SHAKE-128;
- we instantiate H with SHA3-256;
- we instantiate G with SHA3-512;
- we instantiate PRF(s,b) with SHAKE-256$(s||b)$; and
- we instantiate KDF with SHAKE-256.

"90s" variant of Kyber. In the 90s variant of Kyber

- we instantiate XOF(ρ,i,j) with AES-256 in CTR mode, where ρ is used as the key and ikj is zeropadded to a 12-byte nonce. The counter of CTR mode is initialized to zero.
- we instantiate H with SHA-256;
- we instantiate G with SHA-512;
- we instantiate PRF(s,b) with AES-256 in CTR mode, where s is used as the key and b is zero-padded to a 12-byte nonce. The counter of CTR mode is initialized to zero.

- we instantiate KDF with SHA-256.

1.5 Design rationale

The design of Kyber is based on the module version [67] of the Ring-LWE LPR encryption scheme [72] with bit-dropping [84, 90]. It is also enhanced by many of the improvements of preceding implementations of lattice-based encryption schemes such as NewHope [10]. In NewHope (and all other Ring-LWE schemes), operations were of the form $\mathbf{as}+\mathbf{e}$ where all the variables were polynomials in some ring. The main difference in Kyber is that $\mathbf{A}$ is now a matrix (with a small dimension like 3) over a constant-size polynomial ring and $\mathbf{s}$,$\mathbf{e}$ are vectors over the same ring. We refer to this as a scheme over "module lattices."

The use of Module-LWE. Previous proposals of LWE-based cryptosystems either used the very structured Ring-LWE problem (as, for example, NewHope [10]) or standard LWE (as, for example, Frodo [23]). The main advantage of structured LWE variants based on polynomial rings is efficiency in terms of both speed and key and ciphertext sizes. The disadvantages are concerns that the additional structure might enable more efficient attacks and that tradeoffs between efficiency and security can be scaled only rather coarsely. The advantages of standard LWE is the lack of structure and easy scalability, but those come at the cost of significantly decreasing efficiency. Module-LWE offers a trade-off between these two extremes. In the specific case of the Module-LWE parameters used in Kyber, we obtain somewhat reduced structure compared to Ring-LWE, much better scalability, and—when encrypting messages of a fixed size of 256 bits—performance very similar to Ring-LWE-based schemes.

Active security. In [25], Bos, Costello, Naehrig, and Stebila used a *passively secure* KEM to migrate TLS to transitional post-quantum security (i.e., post-quantum confidentiality, but only pre-quantum authentication). Subsequent work, like NewHope [10] or Frodo [23] followed up and proposed more efficient and more conservative instantiations of the underlying passively secure KEM. One advantage of passively secure KEMs is that they can accept a higher failure probability (which allows to either increase security by increasing noise or decreasing public-key and ciphertext size). The other advantage is that they do not require a CCA transform, and therefore come with faster decapsulation. Despite these advantages, Kyber is defined as an IND-CCA2 secure KEM only. For many applications like public-key encryption (via a KEM-DEM construction) or in authenticated key exchange active security is mandatory. However, also in use cases (like key exchange in TLS) that do not strictly speaking require active security, using an actively secure KEM has advantages. Most notably, it allows (intentional or accidental) caching of ephemeral keys. Furthermore, the CCA transform of Kyber protects against certain bugs in implementations. Specifically, passively secure schemes will not notice if the communication partner uses "wrong" noise, for example, all-zero noise. Such a bug in the encapsulation of Kyber will immediately be caught by the re-encryption step during decapsulation. As a conclusion, we believe that the overhead of providing CCA security is not large enough to justify saving it and making the scheme less robust.

The role of the NTT. Multiplication in R_q based on the number-theoretic transform (NTT) has multiple advantages: it is extremely fast, does not require additional memory (like, for example, Karatsuba or Toom multiplication) and can be done in very little code space. Consequently, it has become common practice to choose parameters of lattice-based crypto to support this very fast multiplication algorithm. Some schemes go further and make the NTT part of the definition of the scheme. A prominent example is again NewHope, which samples the public value $\mathbf{a}$ in NTT domain and also sends public keys and ciphertexts in NTT domain to save 2 NTTs. NewHope was not the first scheme to do this; for earlier examples see [70, 90, 96].

In Kyber we also decided to make the NTT part of the definition of the scheme, but only in the sampling of $\mathbf{A}$ and the public key, not for the format of the ciphertext. A consequence of this decision is that the NTT appears in the specification of Kyber.CPAPKE. Note that multiplications by $\mathbf{A}$ have to use the NTT, simply because $\hat{\mathbf{A}}$ is sampled in

NTT domain[2]. Similarly, multiplications by the public-key $\hat{\mathbf{t}}$ have to use the NTT, because the public key is transmitted in NTT domain. As a consequence, implementations will also want to use the NTT for all other multiplications, so we make those invocations of NTT and NTT^{-1} also explicit in Alg. 4, Alg. 5, and Alg. 6. Note that also the secret key sk is stored in NTT domain.

We could have chosen to *not* make the NTT part of the definition of Kyber, which would have increased simplicity of the description. The cost for this increased simplicity would have been k^2 additional NTT operations in both key generation and encapsulation, which would result in a noticeable slowdown. We could also have chosen to not encode the public key in NTT domain; however this would also have required additional NTTs. Finally, we could have chosen to also send the ciphertext in NTT domain; however, this would be incompatible with ciphertext compression via the Compress$_q$ function.

Against all authority. For the generation of the public uniformly random matrix $\mathbf{A}$, we decided to adopt the "against-all-authority" approach of NewHope. This means that the matrix is not a system parameter but instead generated freshly as part of every public key. There are two advantages to this approach: First, this avoids discussions about how exactly a uniformly random system parameter was generated. Second, it protects against the all-for-the-price-of-one attack scenario of an attacker using a serious amount of computation to find a short basis of the lattice spanned by $\mathbf{A}$ *once* and then using this short basis to attack *all* users. The cost for this decision is the expansion of the matrix $\mathbf{A}$ from a random seed during key generation and encapsulation; we discuss this cost more in Subsection 2.1.

Binomial noise. Theoretic treatments of LWE-based encryption typically consider LWE with Gaussian noise, either rounded Gaussian [94] or discrete Gaussian [28]. As a result, many early implementations also sampled noise from a discrete Gaussian distribution, which turns out to be either fairly inefficient (see, for example, [25]) or vulnerable to timing attacks (see, for example, [29, 89, 45]). The performance of the best known attacks against LWE-based encryption does not depend on the exact distribution of noise, but rather on the standard deviation (and potentially the entropy). This motivates the use of noise distributions that we can easily, efficiently, and securely sample from. One example is the centered binomial distribution used in [10]. Another example is the use of "learning-with-rounding" (LWR), which adds deterministic uniform noise by dropping bits as in Kyber's Compress$_q$ function. In the design of Kyber we decided to use centered binomial noise and thus rely on LWE instead of LWR as the underlying problem.

Allowing decapsulation failures. Another interesting design decision is whether to allow decapsulation failures (i.e., decryption failures in Kyber.CPAPKE) or choose parameters that not only have a negligible, but a zero chance of failure. The advantages of zero failure probability are obvious: CCA transforms and security proofs become easier and we can avoid the whole discussion of attacks exploiting decapsulation failures in Subsection 5.5. The disadvantage of designing LWE-based encryption with zero failures is that it means either decreasing security against attacks targeting the underlying lattice problem (by significantly decreasing the noise) or decreasing performance (by compensating for the loss in security via an increase of the lattice dimension). The decision to allow failure probabilities in all parameter sets of Kyber reflects the intuition that

- decapsulation failures are a problem if they appear with non-negligible probability; but
- attacks attempting to exploit failures that occur with extremely low probability as in Kyber are a much smaller threat than, for example, improvements to hybrid attacks [56] targeting schemes with very low noise.

Different noise values η_1 and η_2. Notice that in Algorithm 5, there is additional implicit noise created via Compress$_q$. This has the effect of adding some (deterministic) noise to the ciphertext, which can be interpreted as increasing the

[2] An alternative would be to apply NTT^{-1} to $\hat{\mathbf{A}}$ but that would counteract the whole point of sampling $\mathbf{A}$ in NTT domain.

noise of the error polynomials $\mathbf{e}_1$ and e_2. If the decryption error probability is low enough, then it then makes sense to also increase the noise of the other secret terms (i.e. $\mathbf{s},\mathbf{e},\mathbf{r}$) to be at a similar level as $\mathbf{e}_1$ plus the deterministic noise. We utilize this idea (exclusively) for the Kyber512 parameter set. Relying on the rounding noise from Compress_q to add error is akin to the LWR assumption. But unlike in (Ring/Module)-LWR schemes, where the security completely relies on the noise that's deterministically generated by rounding, our dependence on the deterministic noise is much smaller. First, it only adds 6 bits of Core-SVP hardness, and second, we are adding noise and rounding together, which presumably has less algebraic structure than just rounding. In short, without the LWR assumption, our parameter set for Kyber512 has 112 bits of core-SVP hardness – more specifically, the public keys are protected with 118 bits, and the ciphertexts with 112; with a weak version of the LWR assumption, it has 118-bit security everywhere.

Additional Hashes. In the CCA transform we hash the (hash of the) public key pk into the pre-key $\bar{K}$ and into the random coins r (see line 3 of Alg. 8), and we hash the (hash of the) ciphertext into the final key K. These hashes are not necessary for the security reduction (see Section 4), but they add *robustness*. Specifically, the final shared key output by Kyber.CCAKEM depends on the full view of exchanged messages (public key and ciphertext), which means that the KEM is contributory and safe to use in authenticated key exchanges without additional hashing of context. Hashing pk also into the random coins r adds protection against a certain class of multi-target attacks that attempt to make use of protocol failures. This is discussed in more detail in Subsection 5.5.

Choice of symmetric primitives. In the design of Kyber we need an extendable output function (XOF), two hash functions, a pseudorandom function, and a KDF. We decided to rely on only one underlying primitive for all those functions. This helps to reduce code size in embedded platforms and (for a conservative choice) reduces concerns that Kyber could be attacked by exploiting weaknesses in *one out of several* symmetric primitives. There are only relatively few extendable output functions described in the literature. The best known ones, which also coined the term XOF, are the SHAKE functions based on Keccak [20] and standardized in FIPS-202 [81]. This standard conveniently also describes hash functions with the output lengths we need; furthermore, SHAKE is designed to also work as a PRF. These properties of the FIPS-202 function family made the choice easy, but there are still two decisions that may need explanation:

- We could have chosen to instantiate all symmetric primitives with only *one* function (e.g., SHAKE256) from the FIPS-202 standard. The choice of SHAKE-128 as instantiation of the XOF is actually important for performance; also we do not need any of the traditional security properties of hash functions from SHAKE-128, but rather that the output "looks uniformly random". In an earlier version of Kyber we instantiated H, G, and PRF all with SHAKE-256. We decided to change this to *different* functions from the FIPS-202 family to avoid any domain-separation discussion. Note that this decision increases code-size at most marginally: all 4 functions can be obtained by a call to a "Keccak" function with appropriate arguments (see, for example, [19]).
- We could have decided to use KMAC from NIST Special Publication 800-185 to instantiate the PRF. We decided against this, because it would increase the numbers of Keccak permutations required in the generation of the noise polynomials and thus noticeably and unnecessarily decrease performance.

As a modification in round-2, we decided to derive the final key using SHAKE-256 instead of SHA3-256. This is an advantage for protocols that need keys of more than 256 bits. Instead of first requesting a 256-bit key from Kyber and then expanding it, they can pass an additional key-length parameter to Kyber and obtain a key of the desired length. This feature is not supported by the NIST API, so in our implementations we set the keylength to a fixed length of 32 bytes in api.h.

Choice of symmetric primitives in the "90s" variant The 90s variant of Kyber uses symmetric primitives that are standardized by NIST and accelerated in hardware on a large variety of platforms. These two criteria narrow the choice to AES and SHA-256, which are, for example, implemented in hardware on recent Intel, AMD, and ARM

processors. A natural choice for the hash function G with 512-bit output is SHA-512 from the same SHA-2 family of hash functions as SHA-256.

Supporting non-incremental hash APIs. In line 3 of Alg. 8 we feed H(pk) (instead of pk) into G and in line 5 we feed H(c) (instead of c) into H. Using H(pk) in the call to G enables a small speedup for decapsulation as described in Subsection 2.1. However, there is another reason why we first hash pk and c, namely that it simplifies implementing Kyber with a non-incremental hash API. If Kyber is implemented in an environment which already offers a library for hashing, but only offers calls of the form h = H(m), then producing a hash of the form h = H(m_1km_2) would first require copying m_1 and m_2 into one consecutive area of memory. This would require unnecessary copies and, more importantly, additional stack space. Such non-incremental hash APIs are not uncommon: one example is the API of NaCl [18].

Return value for decapsulation failure. Traditionally the FO transform returns $\perp$ (i.e., a special failure symbol) when decapsulation fails. We use a variant that instead sets the resulting shared key to a pseudorandom value computed as the hash of a secret z and the ciphertext c. This variant of the FO transform was proven secure in [55]. In practice it has the advantage that implementations of Kyber's decapsulation are safe to use even if higher level protocols fail to check the return value. As a consequence of this *implicit rejection* approach, our implementations of decapsulation always return 0.

2 Performance analysis

In this section we consider implementational aspects of Kyber and report performance results of two implementations: the ANSI C reference implementation requested by NIST and an implementation optimized using AVX2 vector instructions. included in the submission package under Additional_Implementations/avx2/. We remark that the optimized implementation in ANSI C in subdirectory Optimized_Implementation/, as requested by the Call for Proposals, is a copy of the reference implementation.

The big picture of Kyber performance. Thanks to the extremely efficient NTT-based multiplication and sampling of $\mathbf{A}$ in NTT domain, the performance of Kyber is largely determined by the performance of the symmetric primitives. This is illustrated, for example, by the fact that the AVX2-optimized implementation of the 90s variant of Kyber is almost twice as fast as Kyber with symmetric primitives based on Keccak. This difference is going to be even larger on systems with hardware-accelerated SHA-256. It is also illustrated by the fact that for optimized implementations, decapsulation is *faster* than encapsulation, even though the former contains a re-encapsulation necessary for the FO transform. The main reason for this is that decapsulation does not need to compute H(pk), which more than compensates for the extra NTT-dominated re-encapsulation.

The fact that (sequential) hashing is very slow in relation to polynomial arithmetic on recent Intel processors also means that the additional hashes discussed in Section 1.5 have a very significant impact on performance: they account for almost half of the encapsulation cycles of our AVX2-optimized implementation on Intel processors.[3] As also mentioned in Section 1.5, hashing the public key and the ciphertext is not required for CCA security, but is instead done to make the function more robust and directly usable in applications that require the shared key to depend on the entire transcript. Our rationale is that because the basic operations comprising Kyber are extremely fast, we can afford to pay a time penalty and make the default version of Kyber as robust and misuse-resilient as possible. Nevertheless, in light of the fact that this hashing has a profound effect on the absolute performance of the scheme, it is an interesting future research direction to see whether there are applications that indeed require hashing the

[3] The reason that the use of Keccak as an XOF to expand the public matrix $\mathbf{A}$ from the seed ρ, and as a PRF to generate secrets and noise, has a smaller impact is because these Keccak invocations are efficiently parallelizable.

entire ciphertext and/or public key. If it's enough to only a hash a small part of them, then the running time of Kyber can be cut in half (on many processors) without sacrificing any robustness.

2.1 Implementation considerations and tradeoffs

Implementing the NTT. Many different tradeoffs are possible when implementing the number-theoretic transform. The most important ones are between code size (which becomes mainly relevant on embedded processors) and speed. The two implementations of Kyber included in the submission package have a dedicated forward NTT (from normal to bitreversed order) and inverse NTT (from bitreversed to normal order). Also, both implementations use precomputed tables of powers of ζ. What is particularly interesting about using the NTT on embedded platforms is that the multiplication of two elements of R_q can be computed without any additional temporary storage. What is particularly interesting about using the NTT on large processors is that it is extremely efficiently vectorizable. Since 2013, the most efficient approach to compute the NTT on 64-bit Intel processors was to represent coefficients as double-precision floating-point values [51, 10]. In our AVX2-optimized implementation of Kyber, we show that carefully optimizing the NTT using AVX2 *integer* instructions results in much better performance. Specifically, on Intel Haswell CPUs one forward NTT in Kyber takes only about 320 cycles; an inverse NTT takes only about 290 cycles.

Keccak. The second speed-critical component inside Kyber are the symmetric primitives, i.e., SHA3-256, SHA3-512, SHAKE-128, and SHAKE-256, all based on the Keccak permutation. SHA3 has the reputation to not be the fastest hash function in software (see, for example, [66]). To some extent this is compensated by the fact that most calls to Keccak are parallel and thus very efficiently vectorizable. Our AVX2 implementation makes use of this fact. Also, ARM recently announced that future ARMv8 processors will have hardware support for SHA3 [50], so there is a good chance that at least on some architectures, software performance of SHA3 will not be an issue in the future.

AES and SHA-2. To illustrate what performance Kyber can achieve with hardware-accelerated symmetric primitives we include the 90s variant using AES, SHA256 and SHA512 instead of symmetric primitives based on Keccak. This variant is interesting only if at least AES is accelerated in hardware, because constant-time software implementations (required for the use of AES as a PRF) are not faster than parallel Keccak.

Hardware-RNGs for key generation. During key generation, the generation of $\mathbf{s}$ and $\mathbf{e}$ is performed using SHAKE-256 (and AES in the 90s variant). However, this is not required. The choice of RNG during key generation is a local decision that any user and platform can make independently. In particular on platforms with fast hardware AES one can adapt the AES-based PRF from the 90s variant also for the otherwise Keccak-based Kyber. We considered using this in our AVX2 implementation, but using this optimization means that testvectors would not match between our two implementations. This is not an issue in actual deployments, where randombytes is not deterministic.

Caching of ephemeral keys. Applications that are even more conscious of key-generation time can decide to cache ephemeral keys for some time. This is enabled by the fact that Kyber is IND-CCA2 secure.

Tradeoffs between secret-key size and speed. It is possible to use different tradeoffs between secret-key size and decapsulation speed. If secret-key size is critical, it is of course possible to not store $\mathrm{H}(pk)$ and also to not store the public key as part of the secret key but instead recompute it during decapsulation. Furthermore, not keeping the secret key in NTT domain makes it possible to compress each coefficient to only 3 bits, which means that a Kyber768 secret key can be compressed into just 288 bytes. This has the additional advantage that it makes key recovery via cold-boot attacks [52] somewhat harder [3]. Finally, as all randomness in key generation is generated from two 32-byte seeds, it is also possible to only store these seeds and re-run key generation during decapsulation. When opting for such 32-byte secret keys, the re-encapsulation step of decapsulation can save the expansion of the matrix $\mathbf{A}$, is it is already expanded (in transposed form) in key generation.

In the other direction, if secret-key size does not matter very much and decapsulation speed is critical, one might decide to store the expanded matrix $\mathbf{A}$ as part of the secret key and avoid recomputation from the seed ρ during the re-encapsulation part of decapsulation.

Both implementations included in the submission package use the secret-key format described in Algorithm 7, i.e., with polynomials in NTT domain, including the public key and H(pk), but not including the matrix $\mathbf{A}$.

Local storage format of static public keys. A user who is frequently encapsulating messages to the same public key can speed up encapsulation by locally storing an expanded public key containing the matrix $\mathbf{A}$ and H(pk). This saves the cost of expanding the matrix $\mathbf{A}$ from the seed ρ and the cost of hashing pk in every encapsulation.

2.2 Performance of Kyber on Intel Haswell CPUs

Table 2 reports Intel Haswell performance results of the reference implementation and an AVX2-optimized implementation of Kyber and the 90s variant of Kyber together with the sizes of keys and ciphertexts. All benchmarks were obtained on one core of an Intel Core i7-4770K (Haswell) processor clocked at 3492MHz (as reported by /proc/cpuinfo) with TurboBoost and hyperthreading disabled. The benchmarking machine has 32GB of RAM and is running Debian GNU/Linux with Linux kernel version 4.19.0. Both implementations were compiled with gcc version 8.3.0 and the compiler flags as indicated in the Makefiles included in the submission package. All cycle counts reported are the median of the cycle counts of 10000 executions of the respective function. The implementations are not optimized for memory usage, but generally Kyber has only very modest memory requirements. This means that in particular our implementations do not need to allocate any memory on the heap.

2.3 Performance of Kyber on ARM Cortex-M4 CPUs

Table 3 reports cycle counts and RAM usage of a C implementation ("clean") and an optimized implementation ("m4") of Kyber on an ARM Cortex-M4. The clean C implementation is a slightly modified version of our C reference implementation from the PQClean project [58]. The optimized implementation is based Table 2: Key and ciphertext sizes and cycle counts for all paramter sets of Kyber. Cycle counts were obtained on one core of an Intel Core i7-4770K (Haswell); "ref" refers to the C reference implementation, "AVX2" to the implementation using AVX2 vector instructions; sk stands for secret key, pk for public key, and ct for ciphertext. In parenthesis are approximate values when including key generation in decapsulation to avoid having to store expanded secret keys. In this scenario, we only store the initial seed d in line 1 of Algorithm 4. The approximate cycle counts for this scenario are computed as the sum of cycle counts for standard decapsulation and key generation minus the number of cycles require to generate the matrix $\mathbf{A}$ from the public seed ρ. Note that this is a very conservative estimate; actual implementations of the approach can also save, for example, sampling the 32 bytes of randomness. See also the discussion on "tradeoffs between secret-key size and speed" in Subsection 2.1.

Kyber512					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	1632 (or 32)	gen:	122684	gen:	33856
pk:	800	enc:	154524	enc:	45200
ct:	768	dec:	187960 (or ≈ 288912)	dec:	34572 (or ≈ 59088)
Kyber512-90s					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	1632 (or 32)	gen:	213156	gen:	21880

pk:	800	enc:	249084	enc:	28592
ct:	768	dec:	277612 (or ≈ 405268)	dec:	20980 (or ≈ 38752)
Kyber768					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	2400 (or 32)	gen:	199408	gen:	52732
pk:	1184	enc:	235260	enc:	67624
ct:	1088	dec:	274900 (or ≈ 425492)	dec:	53156 (or ≈ 82220)
Kyber768-90s					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	2400 (or 32)	gen:	389760	gen:	30460
pk:	1184	enc:	432764	enc:	40140
ct:	1088	dec:	473984 (or ≈ 671864)	dec:	30108 (or ≈ 51512)
Kyber1024					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	3168 (or 32)	gen:	307148	gen:	73544
pk:	1568	enc:	346648	enc:	97324
ct:	1568	dec:	396584 (or ≈ 617848)	dec: 79128 (or ≈ 115332)	
Kyber1024-90s					
Sizes (in Bytes)		Haswell Cycles (ref)		Haswell Cycles (AVX2)	
sk:	3168 (or 32)	gen:	636380	gen:	43212
pk:	1568	enc:	672644	enc:	56556
ct:	1568	dec: 724144 (or ≈ 1009448)		dec:	44328 (or ≈ 71180)

Table 3: Cycle counts and RAM usage (in Bytes) for Kyber on a Cortex-M4 (STM32F407Discovery). The “clean” implementation is a slightly modified C reference implementation from the PQClean project, the “m4” implementation is optimized for the Cortex-M4. All benchmarks are obtained using the pqm4 framework.

Kyber512				
	M4 Cycles (clean)	M4 RAM (clean)	M4 Cycles (m4)	M4 RAM (m4)
gen:	655595	6020	463068	2844
enc:	865256	8668	561518	2484
dec:	961648	9444	519237	2508
Kyber768				
	M4 Cycles (clean)	M4 RAM (clean)	M4 Cycles (m4)	M4 RAM (m4)
gen:	1087897	10052	756224	3292

enc:	1373744	13212	915676	2980
dec:	1491214	14308	853001	3004

Kyber1024

	M4 Cycles (clean)	M4 RAM (clean)	M4 Cycles (m4)	M4 RAM (m4)
gen:	1696314	15180	1213303	3804
enc:	2057522	18844	1407769	3492
dec:	2199958	20420	1326409	3516

on [26] with the improvements from [9] and the round-3 tweaks applied. All benchmarks are obtained using the pqm4 framework [59] on an STM32F407 Discovery board. The two implementations are also included in the submission package. They do not include standalone main functionality, but instead are assumed to be used within pqm4. We do not report benchmarks of the 90s version of Kyber from the Cortex-M4 here, because pqm4 does not yet include a constant-time implementation of AES.

3 Known Answer Test values

All KAT values are included in subdirectories of the directory KAT of the submission package. Specifically, the KAT values of Kyber512 are in the subdirectory KAT/kyber512; the KAT values of the Kyber512-90s are in the subdirectory KAT/kyber512-90s; the KAT values of Kyber768 are in the subdirectory KAT/kyber768; the KAT values of the Kyber768-90s are in the subdirectory KAT/kyber768-90s; the KAT values of Kyber1024 are in the subdirectory KAT/kyber1024; and the KAT values of the Kyber1024-90s are in the subdirectory KAT/kyber1024-90s. Each of those directories contains the KAT values as generated by the PQCgenKAT_kem program provided by NIST. Specifically, those files are:

- KAT/kyber512/PQCkemKAT_1632.req,
- KAT/kyber512/PQCkemKAT_1632.rsp,
- KAT/kyber512-90s/PQCkemKAT_1632.req,
- KAT/kyber512-90s/PQCkemKAT_1632.rsp,
- KAT/kyber768/PQCkemKAT_2400.req,
- KAT/kyber768/PQCkemKAT_2400.rsp,
- KAT/kyber768-90s/PQCkemKAT_2400.req,
- KAT/kyber768-90s/PQCkemKAT_2400.rsp,
- KAT/kyber1024/PQCkemKAT_3168.req,
- KAT/kyber1024/PQCkemKAT_3168.rsp,
- KAT/kyber1024-90s/PQCkemKAT_3168.req, and
- KAT/kyber1024-90s/PQCkemKAT_3168.rsp,

4 Expected security strength

4.1 Security definition

Kyber.CCAKEM (or short, Kyber) is an IND-CCA2-secure key encapsulation mechanism, i.e., it fulfills the security definition stated in Section 4.A.2 of the Call for Proposals.

4.2 Rationale of our security estimates

Our estimates of the security strength for the three different parameter sets of Kyber—and consequently the classification into security levels as defined in Section 4.A.5 of the Call for Proposals—are based on the cost estimates of attacks against the underlying module-learning-with-errors (MLWE) problem as detailed in Subsection 5.1.

To justify this rationale, we will in the following give two reductions from MLWE: a tight reduction in the random-oracle model (ROM) in Theorem 2 and a non-tight reduction in the quantum-random-oracle model (QROM) in Theorem 3. With those reductions at hand, there remain two avenues of attack that would break Kyber without solving the underlying MLWE problem, namely

1. breaking one of the assumptions of the reductions, in particular attacking the symmetric primitives used in Kyber; or
2. exploiting the non-tightness of the QROM reduction.

We briefly discuss 1.) in Subsection 5.4. The discussion of 2.) requires considering two separate issues, namely

- a (quadratic) non-tightness in the decryption-failure probability of Kyber.CPAPKE, and
- a (quadratic) non-tightness between the advantage of the MLWE attacker and the quantum attacker against Kyber.

In Subsection 5.5 we discuss quantum attacks exploiting decryption failures and in the presentation of the non-tight QROM reduction we explain why the non-tightness between quantum attacks against MLWE and quantum attacks against Kyber is unlikely to matter in practice. More specifically, we show how to eliminate this non-tightness if we allow the reasonable, but non-standard, assumption that Kyber.CPAPKE ciphertexts are pseudorandom, even if all randomness is generated pseudorandomly from a hash of the encrypted message.

4.3 Security Assumption

The hard problem underlying the security of our schemes is Module-LWE [27, 67]. It consists in distinguishing uniform samples $(\mathbf{a}_i, b_i) \leftarrow R_q^k \times R_q$ from samples $(\mathbf{a}_i, b_i) \in R_q^k \times R_q$ where $\mathbf{a}_i \leftarrow R_q^k$ is uniform and $b_i = \mathbf{a}_i^T \mathbf{s} + e_i$ with $\mathbf{s} \leftarrow B_\eta^k$ common to all samples and $e_i \leftarrow B_\eta$ fresh for every sample. More precisely, for an algorithm A, we define $\mathbf{Adv}^{\mathrm{mlwe}}_{m,k,\eta}(\mathsf{A}) =$

$$\left| \Pr\left[\mathbf{A} \leftarrow R_q^{m\times k}; (\mathbf{s},\mathbf{e}) \leftarrow \beta_\eta^k \times \beta_\eta^m; b_0 \leftarrow \mathsf{A}(\mathbf{A}, \mathbf{b}) : \mathbf{b} = \mathbf{A}\mathbf{s} + \mathbf{e} \right] - \Pr\left[\mathbf{A} \leftarrow R_q^{m\times k}; \mathbf{b} \leftarrow R_q^m; b_1 \leftarrow \mathsf{A}(\mathbf{A}, \mathbf{b}) : b_1 = 1 \right] \right|$$

4.3.1 Tight reduction from MLWE in the ROM

We first note that Kyber.CPAPKE is tightly IND-CPA secure under the Module-LWE hardness assumption.

Theorem 1. *Suppose* XOF *and G are random oracles. For any adversary* A*, there exist adversaries* B *and* C *with roughly the same running time as that of* A *such that* $\mathbf{Adv}^{\mathrm{cpa}}_{\mathrm{Kyber.CPAPKE}}(\mathsf{A}) \le 2 \cdot \mathbf{Adv}^{\mathrm{mlwe}}_{k+1,k,\eta}(\mathsf{B}) +$

$\mathbf{Adv}_{\mathrm{prfPRF}}(\mathsf{C})$.

The proof of this theorem is easily obtained by noting that, under the MLWE assumption, public-key and ciphertext are pseudo-random.

Kyber.CCAKEM is obtained via a slightly tweaked Fujisaki-Okamoto transform [55, 47] applied to Kyber.CPAPKE. The following concrete security statement proves Kyber.CCAKEM's IND-CCA2-security when the hash functions G and H are modeled as random oracles. It is obtained by combining the generic bounds from [55] with Theorem 1 (and optimizing the constants appearing in the bound).

Theorem 2. *Suppose* XOF*, H, and G are random oracles. For any classical adversary* A *that makes at most* q_{RO} *many queries to random oracles* XOF*, H and G, there exist adversaries* B *and* C *of roughly the same running time as that of* A *such that*

$$\mathbf{Adv}_{\mathrm{ccaKyber.CCAKEM}}(\mathsf{A}) \le 2\mathbf{Adv}_{\mathrm{mlwe}k+1,k,\eta}(\mathsf{B}) + \mathbf{Adv}_{\mathrm{prfPRF}}(\mathsf{C}) + 4q_{RO}\delta.$$

Note that the security bound is tight. The negligible additive term $4q_{RO}\delta$ stems from Kyber.CPAPKE's decryption-failure probability δ.

4.3.2 Non-tight reduction from MLWE in the QROM

As for security in the quantum random oracle model (QROM), [55, 97] proved that Kyber.CCAKEM is INDCCA2 secure in the QROM, provided that Kyber.CPAPKE is IND-CPA secure. A slightly tighter reduction can be obtained by requiring the base scheme Kyber.CPAPKE to be pseudo-random. Pseudo-randomness [97] requires that, for every message m, a (randomly generated) ciphertext $(c_1,c_2) \leftarrow$ Kyber.CPAPKE.Enc(pk,m) is computationally indistinguishable from a random ciphertext of the form (Compress$_q(\mathbf{u},d_u)$,Compress$_q(v,d_v)$),
for uniform $(\mathbf{u},v)$. (We also require the property of "statistical disjointness" [97] which is trivially fulfilled for Kyber.CPAPKE.) The proof of Kyber.CPAPKE's IND-CPA security indeed shows that Kyber.CPAPKE is tightly pseudo-random under the Module-LWE hardness assumption.

Theorem 3. *Suppose* XOF*, H, and G are random oracles. For any quantum adversary* A *that makes at most* q_{RO} *many queries to quantum random oracles* XOF*, H and G, there exist quantum adversaries* B *and* C *of roughly the same running time as that of* A *such that*

$$\mathbf{Adv}_{\mathrm{ccaKyber.CCAKEM}}(\mathsf{A}) \le 4q_{RO} \cdot \sqrt{\mathbf{Adv}^{\mathrm{mlwe}}_{k+1,k,\eta}(\mathsf{B})} + \mathbf{Adv}_{\mathrm{prfPRF}}(\mathsf{C}) + 8q_{RO}^2\delta.$$

Unfortunately, the above security bound is non-tight and therefore can only serve as an asymptotic indication of Kyber.CCAKEM's CCA-security in the quantum random oracle model.

Tight reduction under non-standard assumption. We can use [55, 97] to derive a tight security bound in the QROM from a non-standard security assumption, namely that a deterministic version of Kyber.CPAPKE, called DKyber.CPAPKE, is pseudo-random in the QROM. Deterministic Kyber.CPAPKE is defined as Kyber.CPAPKE, but the random coins r used in encryption are derived deterministically from the message m as $r :=$ G(m). Pseudo-randomness for deterministic encryption states that an encryption (c_1,c_2) of a randomly chosen message is computationally indistinguishable from a random ciphertext (Compress$_q(\mathbf{u},d_u)$,Compress$_q(v,d_v)$), for uniform $(\mathbf{u},v)$. In the classical ROM, pseudo-randomness of DKyber.CPAPKE is tightly equivalent to MLWE but in the QROM the reduction is non-tight (and is the

reason for the term $q_{RO} \cdot \sqrt{\mathbf{Adv}^{\mathrm{mlwe}}_{k+1,k,\eta}(\mathsf{B})}$ in Theorem 3). Concretely, we obtain the following bound:

$$\mathbf{Adv}_{\mathrm{ccaKyber.CCAKEM}}(\mathsf{A}) \leq 2\mathbf{Adv}_{\mathrm{mlwe}k+1,k,\eta}(\mathsf{B}) + \mathbf{Adv}_{\mathrm{DKyberpr.CPAPKE}}(\mathsf{C}) + \mathbf{Adv}_{\mathrm{prfPRF}}(\mathsf{D}) + 8q_{RO2} \qquad \delta.$$

We remark that we are not aware of any quantum attack on deterministic Kyber.CPAPKE that performs better than breaking the MLWE problem.

Table 4: Classical and quantum hardness of the different proposed parameter sets of Kyber together with the claimed security level as defined in Section 4.A.5 of the Call for Proposals.

	Kyber512	Kyber768	Kyber1024
NIST Security level	1	3	5
Core-SVP methodology, Primal attack only (Sec. 5.1.4)			
lattice attack dim. d	999	1419	1885
BKZ-blocksize β	406	626	878
core-SVP classical hardness	118	183	256
core-SVP quantum hardness	107	166	232
Refined estimate for classical attacks using [40, 36, 5] (Sec. 5.2)			
lattice attack dim. d	1025	1467	1918
BKZ-blocksize β	413	637	894
sieving dimension $\beta' = \beta - d_{4f}$	375	586	829
$\log_2$(gates)	151.5	215.1	287.3
$\log_2$(memory in bits)	93.8	138.5	189.7

4.4 Estimated security strength

Table 4 lists the security levels according to the definition in Section 4.A.5 of the Call for Proposals for the different parameter sets of Kyber. Our claims are based on the cost estimates of the best known attacks against the MLWE problem underlying Kyber as detailed in Subsection 5.1. Specifically we list the classical and the quantum *core-SVP hardness* and use those to derive security levels.

All parameter sets of Kyber have some probability of decryption failure. These failures are a security concern (see Section 5.5) and so the probabilities with which they occur need to be small. But because in the classical random oracle model, the decryption failure probability is information-theoretic, we do not see a need for it to decrease with the security parameter. In particular, decryption failure for our level 3 and 5 parameter sets is less than 2^{-160}, which means that if 2^{80} instances of Kyber were run every second from now until our sun becomes a white dwarf, the odds still heavily favor there never being a decryption failure. We therefore exclude these attacks from our claims regarding the NIST security estimates.

The impact of the deterministic noise caused by $\mathrm{Compress}_q$ on Kyber512. Each coefficient of $\mathbf{e}_1$ (and $\mathbf{e}_2$) in Algorithm 5 is distributed as a binomial distribution with parameter $\eta_2 = 2$, which has variance $\eta_2/2 = 1$. The parameter $d_u = 10$ implies that the $\mathrm{Compress}_q$ function maps elements modulo q to a set of size 2^{10}. When this set is mapped back to Z_q via the $\mathrm{Decompress}_q$ function, the difference between every two elements in the latter set is either 3 or 4. This implies that the error created by the $\mathrm{Compress}_q$ function for each coefficient is either uniform over $\{-1,0,1\}$,$\{-1,0,1,2\}$, or $\{-2,-1,0,1\}$. It therefore has variance at least as large as the uniform distribution over the set $\{-1,0,1\}$, which is $2/3$. This makes the total variance of each coefficient of $\mathbf{e}_1$ plus the deterministic error at least $1+2/3 = 5/3$. The other

secret and noise terms have binomial distributions with parameter $\eta_1 = 3$ for a variance of $\eta_1/2 = 3/2 < 5/3$. When accounting for the errors added by Compress$_q$, we therefore calculate the hardness of Kyber512 assuming that it is as hard as Module-LWE where the variance of each secret/error coefficient is $3/2$.

We now make the above discussion more precise. We define the Module-LWER problem as the problem of distinguishing samples $(\mathbf{a}_i, \text{Compress}_q(b_i, d))$ for random $(\mathbf{a}_i, b_i) \leftarrow R_q^k \times R_q$ from samples where $\mathbf{a}_i \leftarrow R_q^k$ and $b_i = \mathbf{a}_i^T \mathbf{s} + e_i$ with $\mathbf{s} \leftarrow B_{\eta_1}^k$ common to all samples and $e_i \leftarrow B_{\eta_2}$ fresh for every sample.[4] Using the same terminology as in the beginning of Section 4.3, we define the advantage of an algorithm A breaking Module-LWER as $\mathbf{Adv}^{\text{mlwer}}_{m,k,\eta_1,\eta_2,d}(\mathsf{A}) =$

$$\Pr[b0 = 1 : \mathbf{b}^{\mathbf{A}} \leftarrow R^{m\times k}; (\mathbf{s}, \mathbf{e}) \leftarrow \beta^k \times \beta^m; \mathbf{b} = \mathbf{As} + \mathbf{e}; b \leftarrow \mathsf{A}(\mathbf{A}, \text{Compress}_q(\mathbf{b}, d))] - \Pr[\mathbf{A} \leftarrow R^{m\times k}; \mathbf{b} \leftarrow R^m; b \leftarrow \mathsf{A}(\mathbf{A}, \text{Compress}_q(\mathbf{b}, d))] \Big|_{\eta_1\eta_2}$$

In addition to the parameters in the Module-LWE problem, the Module-LWER problem has the additional parameter d which defines the range (i.e., $\{0,\ldots,2^d - 1\}$) of the Compress$_q$ function. Also, the η parameter is split into η_1, which defines the distribution of $\mathbf{s}$, and η_2, which defines the distribution of $\mathbf{e}$. The intuition for having $\eta_1 \neq \eta_2$ is that the Compress$_q$ function creates another noise vector $\mathbf{e}^0$ which gets added to $\mathbf{e}$ (we discuss this in more detail below), and so the short target vector crucial for the lattice attack is $\begin{bmatrix} \mathbf{s} \\ \mathbf{e} + \mathbf{e}' \end{bmatrix}$. If we would like to have this vector be balanced, we should set $\eta_1 > \eta_2$.

We can then obtain the analogues of Theorems 1 and 2 with the term $2^{\mathbf{Adv}^{\text{mlwe}}_{k+1,k,\eta}(\mathsf{B})}$ being replaced by

$$\mathbf{Adv}^{\text{mlwe}}_{k,k,\eta_1}(\mathsf{B}) + \mathbf{Adv}^{\text{mlwer}}_{k+1,k,\eta_1,\eta_2,d_u}(\mathsf{B}'), \tag{6}$$

where the Module-LWE assumption is used in the key generation (as before) and the Module-LWER assumption is used in the encryption. We now assess the security of the Module-LWER problem. First, when $\eta_1 \geq \eta_2$, any algorithm B for Module-LWE implies an algorithm A for Module-LWER such that $\mathbf{Adv}^{\text{mlwer}}_{m,k,\eta_1,\eta_2,d}(\mathsf{A}) \leq \mathbf{Adv}^{\text{mlwe}}_{m,k,\eta_2}(\mathsf{B})$. When $\eta_1 > \eta_2$, as in the Kyber512 parameter set, the preceding inequality ignores the fact that having a larger $\mathbf{s}$ in the Module-LWER instance makes the short target lattice vector longer. And when $d < \log q$, the inequality similarly ignores the fact that the adversary A only gets to see $(\mathbf{A}, \mathbf{t} = \text{Compress}_q(\mathbf{As} + \mathbf{e}, d))$ rather than $(\mathbf{A}, \mathbf{As} + \mathbf{e})$. If we use this rough inequality, then for our Kyber512 parameters we would have the Core-SVP classical hardness of the Module-LWER problem be 112. So the logarithm of the required number of gates stated in the Kyber512 column of Table 4 would decrease by 6 to approximately 145.

We argue, however, that based on the current state of knowledge of lattice attacks, the Module-LWER problem with Kyber512 parameters is somewhat harder. In particular, we contend that it is at least as hard as the Module-LWE problem where $\mathbf{s}, \mathbf{e} \leftarrow \eta_1$. This implies that (6)$\leq 2^{\mathbf{Adv}^{\text{mlwe}}_{k+1,k,\eta_1}(\mathsf{B})}$. If we then analyze the hardness of Module-LWE with these parameters, we get the numbers in Table 4.

The quantity that determines the efficacy of the attacks (in terms of the Core-SVP measurement) against the Module-LWE problem is the norm of the vector $\begin{bmatrix} \mathbf{s} \\ \mathbf{e} \end{bmatrix}$ where $\mathbf{As} + \mathbf{e} = \mathbf{b}$ over the ring R_q. In order to create such an equation out of a Module-LWER instance, we map $\mathbf{t} = \text{Compress}_q(\mathbf{As}+\mathbf{e}, d)$ to R_q by applying

[4] This problem generalizes both Module-LWE and Module-LWR. When $e_i = 0$, this problem is Module-LWR, whereas when

Decompress$_q$($\mathbf{t}$,d), which allows us to write $\mathbf{As}+\mathbf{e}+\mathbf{e}^0 = \mathbf{t}$, where $\mathbf{s} \xleftarrow{} \beta_{\eta_1}^k, \mathbf{e} \leftarrow \beta_{\eta_2}^m$ and $\mathbf{e}^0$ is the additional error term caused by the compression. Without knowledge of $\mathbf{s}$ and $\mathbf{e}$, such a mapping is the best that one can use to minimize the norm of $\mathbf{e}^0$. To finish our argument, it would suffice to show that the vector $\mathbf{e} + \mathbf{e}^0$ has larger norm than a vector whose coefficients are drawn from the distribution β_{η_1}. For this, we use the heuristic assumption that $\mathbf{e}^0$ is independent of $\mathbf{s}$ and $\mathbf{e}$ and each of its coefficients has the distribution

$$\text{Decompress}_q \text{Compress}_q(x, d), d) - x \bmod^{\pm} q;\ x \leftarrow \mathbb{Z}_q.^{4} \tag{7}$$

When $d = d_u = 10$ and $q = 3329$ are as in the Kyber512 parameter set, then the distribution in (7) is:

$$\Pr[e' \in \{-2, 2\}] = \frac{257}{3329},\ \Pr[e' \in \{-1, 1\}] = \frac{2048}{3329},\ \Pr[e' = 0] = \frac{1024}{3329}. \tag{8}$$

Using the above distribution on the coefficients of e^0 and the fact that each coefficient of $\mathbf{e}$ comes from the distribution β_{η_2}, we can exactly compute the distribution of the squared norm of $\mathbf{e} + \mathbf{e}^0$. In Table 5, we compare some bounds of this distribution with that of a 512-dimensional vector whose coefficients are all chosen from β_{η_1}. It can be seen from the table that the norm of the vector generated as $\mathbf{e} + \mathbf{e}^0$ is noticeably larger, at all probabilities, than the one generated from β_{η_1}. In fact, because the variance of each coefficient of $\mathbf{e}$ is 1 and of the distribution in (8) (when the probabilities of the ± coefficients are

the compression coefficient d is greater than log$q$, this is the Module-LWE problem.

[4]This assumption (with $\mathbf{e} = 0$) is also used in the security and decryption error analysis of (Module)-LWR problems.

p	2^{-1}	2^{-30}	2^{-60}	2^{-90}	2^{-120}
max α s.t. $\Pr^{X=\mathbf{e}+\mathbf{e}';\mathbf{e}\leftarrow\beta_{\eta_2}^{512},\mathbf{e}'\leftarrow(8)}[\|\|X\|\|^2 < \alpha] < p$	984	679	561	480	419
max α s.t. $\Pr^{X\leftarrow\beta_{\eta_1}^{512}}[\|\|X\|\|^2 < \alpha] < p$	767	529	436	373	325

Table 5: Probability bounds on $\|X\|^2$ when X is distributed as $\mathbf{e} + \mathbf{e}^0$ or as a binomial. For example, when $X = \mathbf{e} + \mathbf{e}^0$, then $\Pr[\|X\|^2 < 561] < 2^{-60}$ while $\Pr[\|X\|^2 < 562] \geq 2^{-60}$.

balanced) is approximately 0.9, the distribution of $\mathbf{e}+\mathbf{e}^0$ behaves very similarly to a Gaussian with variance approximately 1.9. It is therefore reasonable to assume that this error distribution makes the problem harder than if the distribution were β_3, which behaves similarly to a continuous Gaussian of variance 1.5. In short, by also accounting for the rounding error $\mathbf{e}^0$, we gain an additional 6 bits of security with respect to the currently-best attacks.

The impact of MAXDEPTH. The best known quantum speedups for the sieving algorithm, which we consider in our cost analysis (see Subsection 5.1.1), are only mildly affected by limiting the depth of a quantum circuit, because it uses Grover search on sets of small size (compared to searching through the whole keyspace of AES). For the core-SVP-hardness operation estimates to match the quantum gate cost of breaking AES at the respective security levels, a quantum computer would need to support a maximum depth of 70–80. When limiting the maximum depth to smaller values, or when considering classical attacks, the core-SVP-hardness estimates are smaller than the gate counts for attacks against AES. Recent study [5] on the concrete cost of sieving suggest that the quantum speed-ups of these algorithms are tenuous, independently of the value of
MAXDEPTH.

4.5 Additional security properties

4.5.1 Forward secrecy.

Kyber has a very efficient key-generation procedure (see also Section 2) and is therefore particularly well suited for applications that use frequent key generations to achieve forward secrecy.

4.5.2 Side-channel attacks.

Timing attacks. Neither straight-forward reference implementations nor optimized implementations of Kyber use any secret-dependent branches or table lookups[5].This means that typical implementations of Kyber are free from the two most notorious sources of timing leakage. Another possible source of timing leakage are non-constant-time multipliers like the UMULL instruction on ARM Cortex-M3 processors, which multiplies two 32-bit integers to obtain a 64-bit result. However, multiplications in Kyber have only 16-bit inputs, and most non-constant-time multipliers show timing variation only for larger inputs. For example, on ARM Cortex-M3 processors the obvious way to implement multiplications in Kyber is through the constanttime MUL instruction, which multiplies two 32-bit integers, but returns only the bottom 32-bits of the result. What remains as a source of timing leakage are modular reductions, which are sometimes implemented via conditional statements. However, timing leakage in modular reductions is easily avoided by using (faster) Montgomery [79] and Barrett reductions [16] as illustrated in our reference and AVX2 implementations.

We note that the 90s variant is only really attractive to use if AES hardware support is available. If hardware support is not available, then table-based implementations of AES are notorious for leaking secrets trough cache timing. In our C reference implementation of the 90s variant we use a constant-time bitsliced implementation of AES, which is based on code from BearSSL [91].

Differential attacks. We expect that any implementation of Kyber without dedicated protection against differential power or electromagnetic radiation (EM) attacks will be vulnerable to such attacks. This is true for essentially any implementation of a cryptographic scheme that uses long-term (non-ephemeral) keys. Deployment scenarios of Kyber in which an attacker is assumed to have the power to mount such an attack require specially protected—typically masked—implementations. In [83], Oder, Schneider, Pöppelmann, and Güneysu present such a masked implementation of Ring-LWE decryption with a CCA transform very similar to the one used in Kyber. The implementation targets Cortex-M4F microcontrollers; the conclusion of the work is that protecting the decryption (decapsulation) step against first-order DPA incurs an overhead of about about a factor of 5.5. The techniques presented in that paper also apply to Kyber and we expect that the overhead for protecting Kyber against differential attacks is in the same ballpark.

Template attacks. Protections against differential attacks do not help if an attacker is able to recover even ephemeral secrets from a single power or EM trace. At CHES 2017, Primas, Pessl, and Mangard presented such a single-trace attack against an implementation of Ring-LWE on a Cortex-M4F microcontroller [92]. The attacker model in this attack is rather strong: it is the typical setting of template attacks, which assumes an attacker who is able to generate template traces on known inputs on a device with leakage very similar to the actual target device. In [92], the authors used *the same device* for generating target traces and in the attack. The attack was facilitated (maybe even enabled) by the fact that the implementation under attack used variable-time modular reductions. Consequently, the paper states that *"One of the first measures to strengthen an implementation against SPA attacks is to ensure a constant runtime and control flow"*. This is the case for all implementations of Kyber. The attack from [92] would thus certainly not straight-forwardly apply to implementations of Kyber, but more research is required to investigate whether also constant-time implementations of Kyber (and other lattice-based schemes) succumb to template attacks, and what the cost of suitable countermeasures is.

4.5.3 Multi-target attacks

Our security analysis makes no formal claims about security bounds in the multi-target setting. However, in the design of Kyber we made two decisions that aim at improving security against attackers targeting multiple users:

[5] Note that the rejection sampling in generating the matrix $\mathbf{A}$ does not involve any secret data.

- We adopt the "against-all-authority" approach of re-generating the matrix $\mathbf{A}$ for each public key from NewHope [10]. This protects against an attacker attempting to break *many keys at the cost of breaking one key*.

- In the CCA transform (see Alg. 8) we hash the public key into the pre-key $\bar{K}$ and the coins r. Making the coins r dependent of the public key protects against precomputation attacks that attempt to break *one out of many keys*. For details, see Subsection 5.5.

4.5.4 Misuse resilience

The first, and most important, line of defense against misuse is the decision to make IND-CCA2 security non-optional. As discussed in Subsection 1.5, it would have been possible to achieve slightly shorter public keys and ciphertexts, and faster decapsulation, in a CPA-secure variant of Kyber. Using IND-CCA2 security by default makes it safe to use Kyber with static keys and as a consequence also to re-use ephemeral keys for some time. What is *not* safe, is to reuse the same randomness in encapsulation, but that randomness is also not exposed to the outside by the API. The CCA transform has a second effect in terms of robustness: it protects against a broken implementation of the noise sampling. A rather peculiar aspect of LWE-based cryptography is that it will pass typical functional tests even if one communication partner does not add any noise (or by accident samples all-zero noise). The deterministic generation of noise via SHAKE-256 during encapsulation and the re-encryption step during decapsulation will reveal such an implementation mistake immediately.

An additional line of defense against misuse is to hash the public-key into the "pre-key" $\bar{K}$ and thus make sure that the KEM is contributory. Only few protocols require a KEM to be contributory and those protocols can always turn a non-contributory KEM into a contributory one by hashing the public key into the final key. Making this hash part of the KEM design in Kyber ensures that nothing will go wrong on the protocol level if implementers omit the hash there.

A similar statement holds for additionally hashing the ciphertext into the final key. Several protocols need to ensure that the key depends on the complete view of exchanged protocol messages. This is the case, for example, for the authenticated-key-exchange protocols described in the Kyber paper [24, Sec. 5].

Hashing the full protocol view (public key and ciphertext) into the final key already as part of the KEM makes it unnecessary (although of course still safe) to take care of these hashes on the higher protocol layer.

5 Analysis with respect to known attacks

5.1 Attacks against the underlying MLWE problem

MLWE as LWE. The best known attacks against the underlying MLWE problem in Kyber do not make use of the structure in the lattice. We therefore analyze the hardness of the MLWE problem as an LWE problem. We briefly discuss the current state of the art in *algebraic attacks*, i.e., attacks that exploit the structure of module lattices (or ideal lattices) at the end of this subsection.

5.1.1 Attacks against LWE

Many algorithms exist for solving LWE (for a survey see [7]), but many of those are irrelevant for our parameter set. In particular, because there are only

$$m = (k + 1)n$$

LWE samples available to the attacker, we can rule out BKW type of attacks [60] and linearization attacks [13]. This essentially leaves us with two BKZ [98, 32] attacks, usually referred to as primal and dual attacks that we will recall in Subsections 5.1.2 and 5.1.3.

The algorithm BKZ proceeds by reducing a lattice basis using an SVP oracle in a smaller dimension b. It is known [53] that the number of calls to that oracle remains polynomial, yet concretely evaluating the number of calls is rather painful, and this is subject to new heuristic ideas [32, 31, 11].

We start with an analysis that ignores this polynomial factor, i.e. only considers the cost of *a single call* to an SVP oracle in dimension b. We will also use for now a very simple cost estimate for the hardness of SVP. This *core-SVP hardness* methodology was introduced in [10, Sec. 6], as a simple way of estimating security. In the light of the cryptanalytic progress that happened during the NIST evaluation Rounds 1 and 2, we consider that this method remains informative, but too coarse to produce accurate security estimates, especially the for security against classical attackers. This will be addressed by the new Section 5.2, added for the Round 3.

Enumeration vs. sieving. There are two algorithmic approaches for the SVP oracle in BKZ: enumeration and sieving algorithms. These two classes of algorithms have very different performance characteristics and, in particular for sieving, it is hard to predict how *practical performance* scales from lattice dimensions that have been successfully tackled to larger dimensions that are relevant in attacks against cryptosystems like Kyber. The starting point of such an analysis is the fact that enumeration algorithms have super-exponential running time, while sieving algorithms have only exponential running time. Experimental evidence from typical implementations of BKZ [49, 32, 38] shows that enumeration algorithms are more efficient in "small" dimensions, so one question is at what dimension sieving becomes more efficient. At the beginning of the NIST PQC project, the best known sieving techniques were slower in practice for accessible dimensions of up to $b \approx 130$.

The work [40] showed (in the classical setting) that sieving techniques can be sped up *in practice* for exact-SVP, becoming less than an order of magnitude slower than enumeration already in dimension 60 to 80. One reason for the improvement is the "dimensions-for-free" technique, allowing to solve SVP in dimension b by sieving in a sligthly smaller dimension $b' = b - d_{4f}$. Further algorithmic and implementation efforts have finally managed to get sieving to outperform enumeration in practice [4], with a cross-over around dimension 80.

The hardness estimation is complicated by the fact that sieving algorithms are much more *memory intensive* than enumeration algorithms. Specifically, sieving algorithms have exponential complexity not only in time, but also in memory, while enumeration algorithms require only small amounts of memory. In practice, the cost of access to memory increases with the size of memory, which typically only becomes noticeable once the memory requirement exceeds fast local memory (RAM). There is no study, yet, that investigates the algorithmic optimization and practical performance of sieving using slow background storage.

We follow the approach of [10, Sec. 6] to obtain a conservative lower bound on the performance of both sieving and enumeration for the dimensions that are relevant for the cryptanalysis of Kyber. This approach works in the RAM model, i.e., it assumes that access into even exponentially large memory is free.

A lot of recent work has pushed the efficiency of the original lattice sieving algorithms [82, 78], improving the heuristic complexity from $(4/3)^{b+o(b)} \approx 2^{0.415b+o(b)}$ down to $(3/2)^{b+o(b)} \approx 2^{0.292b+o(b)}$ using *locality-sensitive hashing* (LSH) techniques [63, 17]. Without the dimensions-for-free technique, the hidden subexponential factors are typically much greater than 1 in practice [77, 40].

Most of the sieving algorithms have been shown [65, 62] to benefit from Grover's quantum search algorithm, bringing the complexity down to $2^{0.265b+o(b)}$. However, a concrete analysis shows that, in practice, the quantum speed-up remains tenuous [5].

For our first analysis, we will use $2^{0.292b}$ as the classical and $2^{0.265b}$ and the quantum cost estimate of both the primal and dual attacks with block size (dimension) b. We recall those two attacks in the following.

5.1.2 Primal attack.

The primal attack consists of constructing a unique-SVP instance from the LWE problem and solving it using BKZ. We examine how large the block dimension b is required to be for BKZ to find the unique solution. Given the matrix LWE instance $(\mathbf{A},\mathbf{b} = \mathbf{As}+\mathbf{e})$ one builds the lattice $\Lambda = \{\mathbf{x} \in \mathbb{Z}^{m+kn+1} : (\mathbf{A}|\mathbf{I}_m|-\mathbf{b})\mathbf{x} = \mathbf{0} \bmod q\}$ of dimension $d = m+kn+1$, volume q^m, and with a unique-SVP solution $\mathbf{v} = (\mathbf{s},\mathbf{e},1)$ of norm $\lambda \approx \varsigma\sqrt{kn + m}$, where ς is the standard deviation (i.e., the square root of the variance) of the individual secret / error coefficients. Note that the number of used samples m may be chosen between 0 and $(k+1)n$ in our case and we numerically optimize this choice.

Success condition. We model the behavior of BKZ using the geometric series assumption (which is known to be optimistic from the attacker's point of view), that finds a basis whose Gram-Schmidt norms are given by $\|\mathbf{b}_i^\star\| = \delta^{d-2i-1} \cdot \mathrm{Vol}(\Lambda)^{1/d}$, where $\delta = ((\pi b)^{1/b} \cdot b/2\pi e)^{1/2(b-1)}$ [31, 7]. The unique short vector $\mathbf{v}$ will be detected if the projection of $\mathbf{v}$ onto the vector space spanned by the last b Gram-Schmidt vectors is shorter than $\mathbf{b}_{d-b}^\star$. Its projected norm is expected to be $\varsigma\sqrt{b}$, that is, the attack is successful if and only if

$$\varsigma\sqrt{b} \le \delta^{2b-d-1} \cdot q^{m/d}. \tag{9}$$

We note that this analysis introduced in [10] differs and is more conservative than prior works, which were typically based on the hardness of unique-SVP estimates of [48]. The validity of the new analysis has been confirmed by further analysis and experiments in [6].

5.1.3 Dual attack

The dual attack consists of finding a short vector in the dual lattice $\mathbf{w} \in \Lambda^0 = \{(\mathbf{x},\mathbf{y}) \in \mathbb{Z}^m \times \mathbb{Z}^{kn} : \mathbf{A}^t\mathbf{x} = \mathbf{y} \bmod q\}$. Assume we have found a vector $(\mathbf{x},\mathbf{y})$ of length ℓ and compute $z = \mathbf{v}^t \cdot \mathbf{b} = \mathbf{v}^t\mathbf{As} + \mathbf{v}^t\mathbf{e} = \mathbf{w}^t\mathbf{s} + \mathbf{v}^t\mathbf{e} \bmod q$, which is distributed as a Gaussian of standard deviation $\ell\varsigma$ if $(\mathbf{A},\mathbf{b})$ is indeed an LWE sample (otherwise it is uniform $\bmod q$). Those two distributions have maximal variation distance bounded by $\epsilon = 4\exp(-2\pi^2\tau^2)$, where $\tau = \ell\varsigma/q$, that is, given such a vector of length ℓ one has an advantage ϵ against decision-LWE.

The length ℓ of a vector given by the BKZ algorithm is given by $\ell = \|\mathbf{b}_0\|$. Knowing that Λ^0 has dimension $d = m + kn$ and volume q^{kn} we get $\ell = \delta^{d-1}q^{kn/d}$. Therefore, obtaining an ϵ-distinguisher requires running BKZ with block dimension b, where

$$-2\pi^2\tau^2 \ge \ln(\epsilon/4). \tag{10}$$

Note that small advantages ϵ are not relevant since the agreed key is hashed: an attacker needs an advantage of at least $1/2$ to significantly decrease the search space of the agreed key. He must therefore amplify his success probability by building about $1/\epsilon^2$ many such short vectors. Because the sieving algorithms provides $2^{0.2075b}$ vectors, the attack must be repeated at least R times where

$$R = \max(1, 1/(2^{0.2075b}\epsilon^2)).$$

This makes the conservative assumption that all the vectors provided by the Sieve algorithm are as short as the shortest one.

5.1.4 Core-SVP hardness of Kyber

In Table 4 we list the classical and quantum core-SVP-hardness of the three parameter sets of Kyber. The lower bounds of the cost of the primal and dual attack were computed following the approach outlined above using the analysis script Kyber.py that is available online at https://github.com/pq-crystals/securityestimates.

5.2 Beyond core-SVP hardness

At the time the core-SVP hardness measure was introduced by [10], the best implementations of sieving [17, 77] had performance significantly worse than the $2^{.292b}$ CPU cycles proposed as a conservative estimate by this methodology. This was due to substantial polynomial or even sub-exponential overheads hidden in the complexity analysis of $2^{.292b+o(b)}$ given in [17]. Before this [10] proposed the core-SVP approach, security estimates of lattice schemes were typically based on the cost of SVP via enumeration given in [32, 7], leading to much more aggressive parameters. Beyond the cost of SVP-calls, this methodology also introduced a different prediction of when BKZ solves LWE, which was later confirmed [6] and refined [36].

While doubts were expressed to whether sieving would ever outperform the super-exponential, yet practically smaller, costs of enumeration [32] for relevant cryptographic dimensions, significant progress on sieving algorithms [40, 4] has brought down the cross-over point down to dimension about $b = 80$. In fact, the current SVP records are now held by algorithms that employ sieving[6]. This progress mandates a revision and refinement of Kyber security estimates, especially regading classical attacks. In particular, while it was pretty clear from experiments than the costs hidden in the $o(b)$ before those improvement were positive both in practice and asymptotically, the dimensions-for-free technique of [40] offers a sub-exponential speed-up, making it a priori unclear whether the total $o(b)$ term is positive or negative, both asymptotically and concretely.

In summary, while the core-SVP methodology introduced 5 years ago has pushed designers to be more conservative than previously, it now appears this estimation technique is too coarse to produce accurate security estimates. In the following, we give a refined analysis based on the latest developments described in the literature. We complement this with a discussion of all the approximations, simplifications and foreseeable developments that remain to be explored.

We also note that the choice of the gate count metric was recently discussed in the NIST PQC forum mailing list, and in the case of the algorithm of [17], we do not believe this metric to be very realistic. However, it appears that alternative metrics unavoidably involve physical and technological constants (speed of light, density of information, energy efficiency of gates and data transfers). There seems to be no clear consensus on what some those constants would be with current technology, let alone future technologies. These other metrics would also greatly increase the complications involved in the already delicate tuning and analysis of sieving.

In the following subsection we therefore discuss the current understanding of attacks against Kyber in the gate-count metric. We focus the discussion on the concrete case of Kyber512. This preliminary analysis gives a cost of 2^{151} gates, which is a 2^8 factor margin over the targeted security of the 2^{143} gates required for attacks against AES. Our discussion of the 'known unknowns' conclude that this number could be affected by a factor of up to 2^{16} in either direction. While there is a risk to see the security claim drop below the 2^{143} bar in the gate count metric, one should consider the choice of the gate count metric itself as a substantial margin. We do not think that even a drop as large as 2^{16} would be catastrophic, in particular given the massive memory requirements that are ignored in the gate-count metric. Having listed and documented the sources of uncertainties, we hope that many of them can be tackled in the months to come, to narrow the
"confidence interval".

5.2.1 A tentative gate-count estimate accounting for recent progress

For concreteness, we focus this discussion to the case of Kyber512. Let us start by defining the *progressivity overhead* $C = 1/(1 - 2^{-.292}) = 5.46$, that is the limit of ratio between $\sum_{i \le b} 2^{.292i+o(i)}$ and $2^{.292b+o(b)}$ as b grows.

Primal Attack Only. Our first point is that, while the core-SVP hardness methodology suggest that the dual attack is slightly cheaper than the primal one, it is in fact significantly more expensive. Indeed, the analysis of the dual attack

[6] https://www.latticechallenge.org/svp-challenge/

of [10] (repeated above) assumes that one gets exponentially many vectors from sieving in the first block that will be as short as the shortest one. In fact, most of them will have a

length $\sqrt{4/3}$ larger. Furthermore, the assumption that we obtain exponentially many such short vectors is incompatible with some of the latest sieving improvements, specifically the dimensions-for-free techinque from [40]. We note that the analysis from [2], which does not assume those extra short vectors, indeed does predict much larger cost for the dual attack than the primal.

BKZ simulation. The analysis of the BKZ success condition from [10, 6] is based on the so called geometric-series assumption, an assumption that has several inaccuracies; in particular it misses a "tail" phenomenon [99]. We instead rely on the simulator provided as part of the leaky-LWE-estimator of [36]. This simulator uses progressive-BKZ [11, 4], which we believe gives somewhat better performance than fixedblock BKZ. It predicts a median success when reaching blocksize $b = 413$. Success grows by a factor of 1.373 from $b = 412$ to $b = 413$, which is larger than the cost factor of $2^{0.292} \approx 1.224$: aiming at lower blocksize and retrying is not a worthy strategy (it grows even faster for $b < 412$). The overall dimension of the lattice is $n = 1025$.

The cost of progressive BKZ with sieving up to blocksize b is essentially $C \cdot (n-b) \approx 3340$ times the cost of sieving for SVP in dimension b.

Dimensions for free. According to [40], the number of "dimensions for free" is $d_{4f} = \frac{b \ln(4/3)}{\ln(b/(2\pi e))}$, which is $d_{4f} = 37.3$ (or 38, rounding upward). That is, each SVP oracle call in dimension $b = 413$ will require running sieving in dimension $b' = b - d_{4f} = 375$.

We note that [4, Fig. 5] obtains a few more dimensions for free in practice than [40] with two tricks: "on-the-fly lifting" and "pump-down sieves". In practice, this saves a bit of memory ($2^{.5}$), and has no visible effect on time. The trick at hand is really fit for CPUs (as it uses more floating-point arithmetic), and probably less interesting for implementations in dedicated circuits (which favor XOR-popcounts).

Gate cost of sieving. Recent work [5] has proposed an analysis of the cost of sieving with a classical and quantum circuit (with RAM access to data). More specifically, they focus on the analysis of a "AllPairSearch" function, account for exact volumes of spherical caps and wedges, compute explicit gate counts for the innermost loop operations (XOR-popcounts, inner products), and automatically tune parameters to obtain concrete classical and quantum costs. Regarding the best classical algorithm, they conclude with a cost of about $2^{137.4}$ gates for AllPairSearch in dimension 375.[7]

The work in [5] is motivated by the quantum/classical speed-up, therefore it does not consider the required number of calls to AllPairSearch. Naive sieving requires a polynomial number of calls to this routine, however this number of calls appears rather small in practice using progressive sieving [40, 64], and we will assume that it needs to be called only once per dimension during progressive sieving, for a cost of $C \cdot 2^{137.4}$ gates[8]. Final gate count. Overall, we conclude with a gate count of

$$G = (1025 - 413) \cdot C_2 \cdot 2_{137.4} = 2_{151.5}.$$

We can also estimate the memory requirement following the analysis of [5], maybe working under the assumption that each coordinate of a lattice vector used during the sieving can be represented within a byte.

[7] This was interpolated from the datafile cost-estimate-list_decoding-classical.csv from the version of May 2020 of [5], available at https://eprint.iacr.org/eprint-bin/getfile.pl?entry=2019/1161&version=20200520:144757&file=1161.pdf . This datafile can be extracted from the pdf via the linux tool pdfdetach. Another version of that datafile is more easily accessible at https://github.com/jschanck/eprint-2019-1161/blob/main/data/cost-estimate-list_decoding-classical. csv, and differs from the one we used by less than a bit.

[8] This is using the fact inside BKZ, doing a single pump per SVP instance seems sufficient [2], while an isolated SVP instance requires an extra progressive loop, called "work-out" on top of the "pump". That is, an isolated SVP instance should have an extra C factor.

Automatizing the above calculations,[9] we obtain Table 4, summarizing the refined estimates for all three parameter sets.

We also note that a similar refined count of quantum gates seems essentially irrelevant for our security claim: the work of [5] concluded that quantum speed-up of sieving are rather tenuous, while the quantum security target for each category is significantly lower than the classical target.

5.3 Approximations, overheads, and foreseeable improvements

We now propose a list of research direction to refine the best attack or its cost estimation, together with an educated guess of their potential concrete impact. We hope that this list of open problems may motivate the community towards a more informed consensus on the actual cost of lattice attacks, that would apply not only to Kyber, but to all lattice-based candidates in the NIST PQC standardization project.

Q1: Idealized Angles and Lengths. A first approximation made in the concrete analysis of [5] is that all vectors have the same length, and therefore that a pair of vectors leads to a reduction precisely if they have an angle below $\pi/3$. However, both heuristically and experimentally, this is not exactly true. Indeed, the length of the sum $\mathrm{k}x \pm y\mathrm{k}$ may be smaller than $\max(\mathrm{k}x\mathrm{k},\mathrm{k}y\mathrm{k})$ despite an angle slightly larger than $\pi/3$ when $\mathrm{k}x\mathrm{k}$ and $\mathrm{k}y\mathrm{k}$ differ a bit. Furthermore, implementations such as [4] even relax the reduction condition, and perform a reduction $z \leftarrow x \pm y$ where z is neither x nor y, but the current worst vector of the database.

This approximation and relaxation may explain the discrepancy of about a factor of 8 between the number of vectors required in the analysis [5] and the experiments up to dimension $b = 120$ [4]. However, these reducing pairs at higher angles are a bit harder to detect with near-neighbors-search techniques.
Expected influence on memory estimate: A factor of 2^{-3}.
Expected influence on gate-count estimate: A factor between 2^{-3} and 2^{-1}.

Q2: Idealized near-neighbors search. The analysis of [5] assumes that the BDGL [17] sieve behaves as if the buckets are perfectly uniformly distributed over the sphere; however the actual algorithm [17] must resort to some structure in the way these buckets are distributed. The original analysis shows that this structure does not affect the success probability of finding each pair by a factor more than subexponential $2^{O(\sqrt{b})}$ [17]. However, this asymptotic analysis may not be that tight, and indeed, preliminary experiments suggest that this success-probability loss may not be so large in practice.

Furthermore, depending on the parametrization, this algorithm has overheads compared to the idealization of [5]: minimizing memory requirement induces a $2^{O(b/\log(b))}$ overhead on time, which can in principle be traded for an asymptotically similar memory overhead.
Expected influence on gate-count estimate: A factor of up to 2^5. (part of which tradeable for more memory)

Q3: Saturation, collisions, and dimensions for free. Another question that affects the concrete performance of the sieving is the delicate choice of when to stop it. In practice the latest implementations [40, 4] chose to stop the sieving when a fraction $s = 1/2$ of the expected vectors of length $\sqrt{4/3} \cdot \mathrm{GH}$ has been found.[10] This means that the success probability of actually solving SVP with dimensions for free can decrease by a factor s (or maybe a bit less because in practice the sieving outputs about 6 times more vectors, the remaining of which are a sligthly longer than $\sqrt{4/3}\mathrm{GH}$, and also have some probability of lifting to the shortest vector).

[9] Scripts available at https://github.com/lducas/leaky-LWE-Estimator/tree/NIST-round3
[10] Here, GH denote the expected length of the shortest vector according to the Gaussian Heuristic.

On the other hand, stopping the sieving with a larger saturation value s also means that we observe more collisions: a pair does reduces, but result more often in an already known vector, making it useless. This must be compensated by enlarging the database, and/or running more reduction rounds. Preliminary experiments showed that the frequency of collision can be significantly larger than s; this may be explained by repeated reduction of the same pair, and the fact that the database tends to not be uniform over the ball, and biased toward shorter vectors.
Expected influence on gate-count estimate: A factor between 2^{-1} and 2^{2}.

Q4: Number of calls to AllPairSearch. The number of calls to AllPairSearch at each step of Progressive Sieving was assumed to be 1. This could be tested experimentally. It is not unthinkable that this is in fact a fractional number less than 1, as, in progressive-sieving, the previous calls in lower dimension may already provide a close to target set of vectors. Preliminary experiments suggest $1/3$ or $1/2$ in very small dimensions; these experiments have been carried without NNS to avoid interference, and deserve to be extended. Expected influence on gate-count estimate: A factor between 2^{-2} and 2^{2}.

Q5: Routing and Congestion (from RAM to circuit). As far as we understand, the sieving circuit studied in [5] still considers RAM access, which, even in the gate count model, and without considering speed-of-light delays easily underestimates cost. For example, once it has been determined in which bucket a vectors belongs, there must be some routing logic to send it there; if there are 2^{x} addresses, then the sent bits must traverse at least $O(x)$ gates. This gets quite more involved as several cores may want to send data to the same bucket at the same time, inducing overheads for queuing logic and/or packet drops. Extra logistic costs are also required in the bucket itself, and if the targeted bucket size is small, variations in the actual numbers of vectors that reach a bucket may induce further overheads or packet drops.

Such logistic costs can be partially mitigated by re-tuning the algorithm: if, once bucketed, data is used often enough, these overheads will become negligible. However, making buckets larger negatively impacts algorithmic performance in the RAM model, and despite mitigation, some loss seems unavoidable. We note that the current parametrization from [5] suggests to target an average bucket size of 32 in the relevant dimensions, leading to each bit of data being used in 32 many XOR-popcounts. We do not think that this parametrization makes logistic overheads negligible.
Expected influence on gate-count estimate: A factor between 2^{2} and 2^{8}.

Q6: Beyond the gate-count metric. While we mostly stick to the gate-count metric in this discussion, we note that the previous research question Q5 might be naturally paired with a similar study in other metrics, maybe having physical and technological constants as parameters. In particular, we would like to mention an ongoing implementation of Sieving with GPU [41], for which optimal performances are reached for quite large bucket-size (say 2^{15}), and this despite using a single computation server (i.e., one is limited by RAM to GPU-RAM bandwidth, and not network bandwidth).

Q7: Refined BKZ strategies. We note that there are ways to refine a bit the (progressive) BKZ strategy used in the simulator of [36]. Indeed, a standard idea [68] is to run a single SVP call in a dimension that is considerably larger than the BKZ blocksize after the BKZ preprocessing, in the hope to partially amortize the $(n-b)\cdot C$ factor. While this has been used for sieving in practice [4] on LWE challenges, a precise analysis of the gains in large dimensions is still missing. We note that this is a time-memory trade-off, as this last SVP call will require more memory than the BKZ pre-processing.

Another folklore idea which remains to be documented and quantified, is to try to swap the “heads” and “tails” of the BKZ profile [99], by instead running dual-BKZ as the pre-processing.

One could also tweak the progressive steps of the blocksize in progressive-BKZ, but our preliminary experiments with the simulations of [36] was not suggesting that significant gains are to be expected there. Expected influence on gate-count estimate: A factor between 2^{-8} and 2^{-2}.

Q8: Module-BKZ in practice. It is known that sieving in cyclotomic/cyclic *ideal* lattices can benefit from speed-ups and memory gain, namely a factor d can be saved on memory, and a factor between $O(d)$ and $O(d^2/\log d)$ can be gained depending on whether Near Neighboor Search techniques are used or not. Namely, the symmetries of those lattices allows to get of full orbit of d vectors of the same length for the price of 1.

In principle, the same gains apply to module lattices of dimension $b = kd$, growing with the module dimension d. It was mentioned on the NIST pqc-forum[11] that a module variant of BKZ could exists, and therefore that its internal SVP oracle could benefit from symmetries. This is confirmed by the theoretical work of [80].

However, there are numerous issues that are likely to prevent substantial gain from this approach. The first one is that one is now very constrained in terms of blocksize: to run module-BKZ with blocksize b on Kyber, one must rely on a sub-module structure of dimension $d = \gcd(b, 256)$. This may force one to over or undershoot the optimal block-size for the attack, and leaves less flexibility, especially for progressivity both at the sieving and BKZ levels. This is even more constraining in combination with the dimension for free, where, as far we can see, one would require both the sieving and the lift context to be aligned with the module structure: $d = \gcd(b, d_{4f}, 256)$. It is also not clear whether, for the same block-size, module-BKZ would give a basis perfectly as good as a BKZ, given that the basis is now constrained by the module structure. One will also need to adjust Kannan embedding to the module structure, which increases the overall dimension of the lattice by d rather than by 1.
Expected influence on gate-count estimate: A factor between 2^{-4} and at most 2^0.

Summary. Combining all the discussion above, it appears that further refinements of the analysis of various aspect of sieving, as well as some foreseeable algorithmic improvements, the estimates may move by a factor somewhere between 2^{-16} and 2^{14}. Again, we hope that at least some of these research directions will soon be documented and elucidated, narrowing down this confidence interval. We recall that they essentially affect all lattice-based candidates in similar way (except maybe Q8).

5.3.1 Algebraic attacks.

While the best known attacks against the MLWE instance underlying Kyber do not make use of the structure in the lattice, we still discuss the current state of the art of such attacks. Most noticeably, several recent works propose new quantum algorithms against Ideal-SVP [44, 30, 21, 34, 35], i.e., solving the shortest vector problem in ideal lattices. The work of [35] mentions obstacles towards a quantum attack on Ring-LWE from their new techniques, but nevertheless suggests using Module-LWE, as it plausibly creates even more obstacles. In [1], Albrecht and Deo establish a reduction from MLWE to RLWE, whose implication is that a polynomial-time algorithm against RLWE with certain parameters would translate to a polynomial-time algorithm against MLWE. In practical terms, however, this attack has a significant slow-down (and this is not just due to the proof) as the dimension of the module increases. This does suggest that increasing the dimension of the module may make the scheme more secure in concrete terms. In particular, going through this reduction to attack Kyber768 would lead to an RLWE problem with quite large modulus and error ($q^0 = q^3$, $\varsigma^0 > q^2\varsigma$), and therefore require the attacker to consider more than 1 sample: the underlying lattice remains a module with a rank strictly larger than 2.

5.4 Attacks against symmetric primitives

All symmetric building blocks of Kyber are instantiated with functions derived from Keccak [20]. In the deterministic expansion of $\mathbf{A}$ from ρ we essentially need SHAKE-128 to produce output that "looks uniformly random" and does not create any backdoors in the underlying lattice problem. In the noise generation we require that concatenating a secret and a public input and feeding this concatenation to SHAKE-256 as input results in a secure pseudorandom

[11] https://groups.google.com/a/list.nist.gov/g/pqc-forum/c/JXN9NWGt9Ys/m/37tgAlU7DAAJ

function. Breaking any of these properties of SHAKE would be a major breakthrough in the cryptanalysis of SHAKE, which would require replacing SHAKE inside Kyber by another XOF.

The security proofs model SHAKE-128, SHA3-256, and SHA3-512 as random oracles, i.e., they are subject to the standard limitations of proofs in the (quantum-)random-oracle model. Turning these limitations into an attack exploiting the instantiation of XOF, H, or G with SHAKE and SHA3 would again constitute a major breakthrough in the understanding of either Keccak or random-oracle proofs in general.

5.5 Attacks exploiting decryption failures

In Theorems 2 and 3 we see that decryption failure probability plays a role in the attacker's advantage: in the classical context in the term $4q_{RO}\delta$ and in the quantum context in the term $8q_{RO}{}^2\,\delta$, where q_{RO} is the number of queries to the (classical or quantum) random oracle.

Attacks exploiting failures. This term in the attacker's advantage is not merely a proof artifact, it can be explained by the following attack: An attacker searches through many different values of m (see line 1 of Alg. 8) until he finds one that produces random coins r (line 3 of Alg. 8) that lead to a decapsulation failure, which will give the attacker information about the secret key. In the quantum setting the search through different values of m is accelerated by Grover's algorithm, which explains the square in the term $q_{RO}{}^2$. With this attack in mind note that with 2^{64} ciphertexts (cmp. Section 4.A.2 of the Call for Proposals), there is a chance of 2^{-100} of a decapsulation failure in Kyber768 without any particular effort by the attacker.

To understand what exactly this means for attacks against Kyber, we need to address the following two questions:

1. How hard is it for an attacker to trigger a Kyber decapsulation failure?
2. How hard is it for an attacker equipped with a ciphertext triggering a failure to mount a successful attack against Kyber?

Regarding the first question, the naive approach of an attacker is to try random ciphertexts, which has a success probability of $q_d\delta$, where q_d is the number of decapsulation queries. In the classical randomoracle model, the cost of an attack exploiting failures will also never get lower than that, as it matches the information-theoretic success probability.

Failure boosting using Grover.

A quantum attacker can try to use Grover search to precompute values of m that have a slightly higher chance to produce a failure. The efficacy of Grover search is limited by the fact that an attacker cannot determine offline whether a given value of m, or more specifically, the derived values $\mathbf{r}$ (line 9 of Alg. 5) and $\mathbf{e}_1$ (line 13 of Alg. 5), produce a decapsulation failure. The reason is that the probability of decapsulation failures largely depends on the products $\mathbf{s}^T\mathbf{e}_1$ and $\mathbf{e}^T\mathbf{r}$ and the attacker does not know (the signs of) the coefficients of $\mathbf{s}$ and $\mathbf{e}$.

Original Kyber analysis. The original Kyber submission document gave an analysis of a particular strategy for using Grover's algorithm to search for values of m that produce $\mathbf{e}_1$ and $\mathbf{r}$ with above-average norm. Intuitively, the larger these values are, the bigger the probability of a decryption failure. The gain achieved through such an approach is, however, limited due to the fact that the distribution of a high-dimensional Gaussian is tightly concentrated around its expected value, while that of a 1-dimensional Gaussian is not as tightly concentrated around its mean. We present this original analysis below, focusing on Kyber768, which has failure probability of 2^{-164}, for concreteness:

The polynomial pair $(\mathbf{e}_1,\mathbf{r})$ can be seen as a vector in $\mathbb{Z}^{1536}$ distributed as a discrete Gaussian with standard deviation $\sigma = \sqrt{\eta_1/2} = 1$ (because $\eta_1 = \eta_2$). By standard tail bounds on discrete Gaussians [14], we know that an m-dimensional vector $\mathbf{v}$ drawn from a discrete Gaussian of standard deviation σ will satisfy

$$\Pr[\|\mathbf{v}\| > \kappa\sigma\sqrt{m}] < \kappa^m \cdot e^{\frac{m}{2}(1-\kappa^2)}, \quad (11)$$

for any $\kappa > 1$.

So for example, the probability of finding a vector which is of length $1.33 \cdot \sigma\sqrt{1536}$ is already as small as 2^{-220}. Even if Grover's algorithm reduces the search space and increases the probability to 2^{-110}, finding such a vector merely increases the chances of getting a decryption error; and the probability increase is governed by the tail-bounds for 1-dimensional Gaussians[12]. For any vector $\mathbf{v}$, if $\mathbf{z}$ is chosen according to a Gaussian with standard deviation σ, then for any κ,

$$\Pr[|\langle \mathbf{z}, \mathbf{v} \rangle| > \kappa\sigma\|\mathbf{v}\|] \leq 2e^{-\kappa^2/2}. \quad (12)$$

If originally, the above probability is set so that decryption errors occur with probability $\approx 2^{-160}$, then $\kappa \approx 15$.[13] If the adversary is then able to increase $\|\mathbf{v}\|$ by a factor of 1.33 (by being able to find larger $(\mathbf{e}_1, \mathbf{r})$), then we can decrease κ by a factor of 1.33 to ≈ 11.25 in (12), which would still give us a probability of a decryption error of less than 2^{-90}. However, finding such a large $\mathbf{v}$ would take at least 2^{110} time, which would make the whole attack cost at least 2^{200}.

Of course one can try to find a slightly smaller $\mathbf{v}$ in the first step so that the entire attack takes less time. If Grover's algorithm really saves a square-root factor, then the optimal value is ≈ 1.05 for κ in (11), which would allow us to lower κ by a factor of 1.05 in (12) to $15/1.05 \approx 14.28$, and would still give a total time to find one decryption error $\approx 2^{-150}$. This makes the attack completely impractical.

Different approaches since 2017. Since the original analysis of failure boosting against Kyber, multiple works have considered multiple aspects of this topic [37, 42]. The latest work on this topic is [22], considers using Grover's algorithm to produce a strategic set of decryption query points which results in a higher chance of triggering a decryption failure *when the number of decryption queries is limited, as in the NIST Call for Proposals.* The overall running time is, however, not less than in the above attack, which didn't restrict itself with a limit on the number of decapsulation queries.

From one to multiple failures. The original Kyber submission document also discussed that an attacker may need more than a single failure to mount a key-recovery attack. For example, the work of [46] required up to 4000 failures, which may suggest that the number of online queries required for a successful attack against Kyber increases by a factor of $\approx 2^{12}$. However, recent work [43] proposed an adaptive strategy, which uses previous failures to significantly lower the cost of the next one. This *directional failure boosting* technique reduces the cost of the full attack to barely more than the cost of triggering the first failure. This means that the answer to the second question in the beginning of this subsection is that one should be on the safe side and make sure that it's hard to trigger even one failure.

Multitarget attacks using failures. Despite the limited gain, an attacker could consider using Grover's algorithm to precompute values of m that produce $\mathbf{r}$ and $\mathbf{e}_1$ with large norm and then use this precomputed set of values of m against many users. This multi-target attack is prevented by hashing the public key pk into the random coins r and thereby into $\mathbf{r}$ and $\mathbf{e}_1$ (line 3 of Alg. 8).

[12] The decryption noise is generated as an inner product of two vectors, and the distribution of this inner product closely resembles the Gaussian distribution.

[13] The above formula only roughly approximates how the decryption error is calculated where $\mathbf{z}$ corresponds to the secret key $(\mathbf{s}, \mathbf{e})$. We should also point out that a part of the decryption error in Kyber is caused by the rounding function Compress, which the adversary has no control over. Therefore this attack will be even less practical than what we describe.

6 Advantages and limitations

6.1 Advantages

In addition to the very competitive speeds, small parameters, and being based on a well-studied problem, the unique advantages of Kyber are:

Ease of implementation: Optimized implementations only have to focus on a fast dimension-256 NTT and a fast Keccak permutation. This will give very competitive performance *for all parameter sets of* Kyber.

Scalability: Switching from one Kyber parameter set to another only requires changing the matrix dimension (i.e., a #define in most C implementations), the noise sampling, and the rounding of the ciphertext via different parameters to the Compress_q function.

We will now give a brief comparison of Kyber to other types of post-quantum schemes (that we are aware of) and, more importantly, to other manners in which lattice-based schemes could be instantiated.

6.2 Comparison to SIDH

An interesting alternative to lattice-based KEMs is supersingular-isogeny Diffie-Hellman (SIDH) [57]. The obvious advantage of SIDH is the sizes of public keys and ciphertexts that—with suitable compression [33]— are about a factor of 3 smaller than Kyber's public keys and ciphertexts. The downside of SIDH is that it is more than 2 orders of magnitude slower than Kyber. The scheme is also rather new, which makes it hard to make definitive comparisons. In the coming years, both implementation speeds and (quantum) attacks against SIDH can improve which may result in faster schemes and/or larger parameters.

6.3 Comparison to code-based KEMs

When considering code-based KEMs, one needs to distinguish the "classical" McEliece and Niederreiter schemes based on binary Goppa codes, and schemes with a less conservative (but more efficient) choice of code. A KEM based on binary Goppa codes can reasonably claim to be a very conservative choice of post-quantum primitive; however, its deployment will, in many scenarios, be hampered by massive publickey size and key-generation time. Less conservative choices, like quasi-cyclic medium-density parity-check (QC-MDPC) codes, are a closer competition in terms of performance but suffer from the fact that for efficient parameters at high security levels they do not achieve (provably) negligible failure probability, which precludes their use in CCA-secure KEMs.

6.4 Comparison to other lattice-based schemes

There are certain design choices that one can make when designing lattice-based schemes, some of which can have significant effects on the efficiency of the resulting scheme and on the underlying security assumption.

Below we list the most important ones and explain the advantages / disadvantages of them versus what we chose for Kyber.

6.4.1 Schemes that build a KEM directly

The Kyber KEM is constructed by encrypting a random message using the LPR encryption [72] (with "bit-dropping"). Another approach one could take is directly building a KEM using the slightly different ideas described in [39, 86]. The advantage of the constructions in [39, 86] over our approach is that if one were to construct a CPA-secure KEM transmitting a b-bit key, then the ciphertext would be b bits shorter, which is about a 3% saving for typical parameters [69]. If, however, one wishes to construct a CCA-secure KEM like Kyber, then this advantage disappears since

transformations from CPA-secure KEMs to CCAsecure ones implicitly go through a CPA-secure encryption scheme, which will result in adding b bits to the KEM. This is why, in Kyber, we simply use the LPR encryption scheme (instead of the CPA-secure key encapsulation) to define Kyber.CPAPKE, and then use this as a building block to construct the IND-CCA2secure KEM Kyber.CCAKEM. Since there is virtually no difference between the two approaches, we will not draw a distinction between schemes constructed in either manner throughout the rest of this section.

6.4.2 LWE based schemes

If one does not want to use any algebraic structure in the LWE problem (i.e. if one takes the MLWE problem over the ring Z), then there are two possibilities for constructing encryption or key-exchange schemes. The first approach makes the public key and the secret key very large (on the order of Megabytes), while keeping the ciphertext at essentially the same size as in Kyber. This type of scheme is the [88] version of the original Regev scheme from [94]. Because of the very large public-key size, this scheme would be extremely inefficient as a key exchange. A scheme more amenable to key exchange is [23], whose public key and ciphertext sizes are both approximately 11KB each, which is approximately 10 times larger than in Kyber. The running time of each party is also larger by a factor of at least 10. In short, LWE-based schemes do not have any ring structure but are an order of magnitude slower and larger than Kyber. They are good back-up schemes in case algebraic structure in lattice schemes could somehow be devastatingly exploited by attackers.

6.4.3 Ring-LWE based schemes

The other extreme in the LWE design space are Ring-LWE (RLWE) schemes based on [72] (e.g., [10]). RLWE is a special case of the MLWE problem where the width of the matrix $\mathbf{A}$ over the ring R is always 1 (and typically, its height would be 2 for a PKE or KEM scheme). Varying the hardness of an RLWE scheme therefore requires to change the dimension of the ring, whereas in Kyber, the ring is always the same and the dimension of the module is being varied. As we mentioned above, one advantage of the approach we chose for Kyber is that we only need to have one good implementation for operations over the ring; varying the dimension of the module simply involves doing more (or fewer) of the same ring operations. Changing the ring, on the other hand, would require completely re-implementing all the operations.

Another advantage of working with a constant-degree "small" ring is that it enables more fine-grained tradeoffs between performance and security. The simplest and most efficient way of implementing RLWE is to work over rings $\mathrm{Z}[X]/(X^n + 1)$ where n is a power of 2. Since n is the only parameter that determines the efficiency and security of RLWE schemes, limiting it to powers of 2 may require overshooting the needed security bound. For example, the dimension of Kyber768 is not reachable. One could of course work directly modulo a polynomial of any desired degree (with the main restriction being that it has to be irreducible over Z), but then the security would decrease slightly due the geometry of non-power-of-2 number fields (see [75, 76]).

The one advantage of RLWE over Kyber is that if $\mathbf{A}$ is a $k \times k$ matrix, then extracting it from a seed requires k times more XOF output than for a 1×1 matrix.

6.4.4 NTRU

When compared to Kyber, NTRU [54] has all the advantages and disadvantages of RLWE, but in addition has two further negative points against it. First NTRU key generation is considerably more expensive than in RLWE when the ring does not support NTT. The reason is that NTRU key generation requires polynomial division, whereas RLWE key generation requires only multiplication (if the ring supports NTT, then division is not much slower than multiplication). The second possible downside of NTRU is that the geometry of its underlying lattice leads to attacks that do not exist against RLWE or MLWE schemes [61]. While this property does not seem to aid in attacks against the small parameters that are used for defining NTRU cryptosystems, it may point to a possible weakness that could be further exploited. The one possible advantage of using NTRU is a small performance advantage during encryption (encapsulation), but

given the disadvantages we do not consider this a good tradeoff. Furthermore, it is not possible to define an efficient version of "Module-NTRU" that would allow for the advantages of Kyber described above in Section 6.1.

6.4.5 Different Polynomial Rings

One could consider using Kyber with a ring that is not $\mathbb{Z}[X]/(X^n + 1)$. An argument that could be made for using different rings is that the rings currently used in Kyber have algebraic properties (e.g., subrings, large Galois groups, etc.) which may be exploited in attacks. We choose to work with $\mathbb{Z}[X]/(X^n + 1)$ for the following reasons:

- From a performance perspective, there is no serious competition; the NTT-based multiplication supported by the parameters we chose for Kyber is at the same time very memory efficient and faster than any other algorithm for multiplication in polynomial rings.
- Lattice-based schemes using the ring $\mathbb{Z}[X]/(X^n + 1)$ have been studied since at least [70]. When the noise vectors are chosen as specified in [72], there have been no improved attacks against RLWE (or MLWE) that use the underlying algebraic structure [87]. Furthermore, being based on MLWE, the algebraic structure of Kyber is very different from that which was exploited in the attacks against *ideal* lattices in [21, 34, 35][14] – we emphasize that the lattice problems underlying the hardness of Kyber are *not* ideal lattices.
- Some of the additional algebraic structure of $\mathbb{Z}[X]/(X^n + 1)$ is actually *helpful against* certain possible attack vectors. As a simple example, it can be proven that when $X^n + 1$ (almost) fully splits modulo q, there do not exist polynomials in the ring that have small norm and many zeros in the NTT representation—the existence of such polynomials for *any* q would weaken the security of MLWE. Along the same lines, it can also be shown that $X^n + 1$ does not have any factors (modulo q) of small degree *and* small norm. For example, all (of which there are exponentially-many in n) factors
  $$\sqrt{}$$
  of degree $n/2$ have ℓ_2-norm at least q. This prevents attacks that map the MLWE instance to a lower-dimensional sub-ring without increasing the errors by too much.
- Finally, $\mathbb{Z}[X]/(X^n + 1)$ is one of the most widely studied, and best understood, rings (along with other cyclotomic rings) in algebraic number theory. The fact that no attacks have been found against its use for cryptosystems like Kyber makes it a much more conservative choice than some ring that is harder to analyze and may show weaknesses only after many more years of study.

6.4.6 Deterministic Noise.

Instead of adding noise $\mathbf{e}, \mathbf{e}_1$, and e_2, one can add "deterministic" noise by simply dropping bits. This is the basis behind the "Learning with Rounding" (LWR) problem [15], which for certain parameters is as hard as the LWE problem. We believe that asymptotically this is a sound approach, but the number of bits that can be dropped before significant decryption error is introduced is not very large (≈ 2 or 3). Relying on just this assumption adds algebraic structure and may allow for a possibility of slightly improved attacks against the scheme. For this reason, and the fact that generating noise is not a particularly costly operation, we did not choose to potentially weaken the scheme to save a little time or have slightly shorter outputs.

7 Brief discussion of relevant results since Nov. 2017

In this section we briefly reference and comment on results relevant to Kyber that were published after the original NIST PQC submission deadline in November 2017.

[14] Also, like the attacks against NTRU, these do not apply for the small parameters used public key encryption schemes.

"small Kyber". The idea of decreasing the modulus q from $q = 7681$ to $q = 3329$ for Kyber was already present in [100]; the paper refers to the resulting parameter set as "small-Kyber". Other parameters chosen in small-Kyber are different than what we describe in this round-2 update; in particular, small-Kyber does not eliminate the public-key compression. More importantly, the polynomial multiplication using the NTT algorithm is different in small-Kyber. Our multiplication algorithm uses the NTT algorithm to compute the decomposition of polynomials $a_i \in Z_q[X]/(X^{256} + 1)$ as

$$(a_i \bmod X^2 - r_1, \ldots, a_i \bmod X^2 - r_{128}),$$

and multiplication is performed using pointwise products modulo $X^2 - r_j$. The algorithm in [100], on the other hand, divides the a_i into two polynomials of degree 128 and performs a "full-splitting" NTT (i.e. modulo $X - r'_i$) on each of them. This method, which computes two (smaller) NTTs and some additional steps to reassemble the polynomials, has slightly worse performance than the original Kyber and considerably worse than Kyber with our round-2 tweaks; though we believe that applying our new tweaks would speed up the algorithm in [100] as well.

Fault attacks. In [93], Ravi, Roy, Bhasin, Chattopadhyay, and Mukhopadhyay describe a fault attack against implementations of NewHope, Kyber, Frodo, and Dilithium on an ARM Cortex-M4. Specifically, these attacks target the instructions loading the nonce in the PRF; skipping this load through a clock glitch reuses the same nonce multiple times, which leads to efficiently breakable instances of the schemes. We could have decided to modify the *specification* of Kyber to generate randomness in a different way. Indeed, generating all randomness at once through one call to the PRF would thwart this specific attack and *maybe* aid future fault-attack protected implementations. However, such a change to the specification would reduce parallelism and thus cost performance on many platforms. More importantly however, the benefits of such a change are still completely unclear. It is not at all surprising that an implementation without any faultattack countermeasures, such as the one targeted in [93], succumbs to *some* fault attack. It is not clear if the attack described in [93] is the most powerful attack and it is also unclear what serious fault-attack countermeasures for any of the targeted schemes have to look like to thwart *any* realistic attack.

Improvements to sieving. Various papers have improved our understanding of attacks against latticebased cryptosystems and in particular the performance of sieving algorithms. Instead of listing those papers here, we updated the discussion in Sections 5.1 and 5.2.

Attacks exploiting decapsulation failures. Since the beginning of the NIST PQC project, several papers investigate chosen-ciphertext attacks exploiting decapsulation failures. Instead of listing those papers here, we updated the discussion in Section 5.5.

References

[1] Martin Albrecht and Amit Deo. Large modulus Ring-LWE ≥ Module-LWE, 2017. To appear. https://eprint.iacr.org/2017/612. 31

[2] Martin R Albrecht, Benjamin R Curtis, Amit Deo, Alex Davidson, Rachel Player, Eamonn W Postlethwaite, Fernando Virdia, and Thomas Wunderer. Estimate all the {LWE, NTRU} schemes! In *International Conference on Security and Cryptography for Networks*, pages 351–367. Springer, 2018. 28

[3] Martin R. Albrecht, Amit Deo, and Kenneth G. Paterson. Cold boot attacks on ring and module LWE keys under the NTT. *Transactions on Cryptographic Hardware and Embedded Systems*, 2018(3):173– 213, 2018. https://doi.org/10.13154/tches.v2018.i3.173-213. 16

[4] Martin R Albrecht, Léo Ducas, Gottfried Herold, Elena Kirshanova, Eamonn W Postlethwaite, and Marc Stevens. The general sieve kernel and new records in lattice reduction. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 717–746. Springer, 2019. 25, 27, 28, 29, 30

[5] Martin R Albrecht, Vlad Gheorghiu, Eamonn W Postlethwaite, and John M Schanck. Estimating quantum speedups for lattice sieves. Technical report, Cryptology ePrint Archive, Report 2019/1161, 2019. 21, 23, 26, 28, 29, 30

[6] Martin R Albrecht, Florian Göpfert, Fernando Virdia, and Thomas Wunderer. Revisiting the expected cost of solving uSVP and applications to LWE. In Tsuyoshi Takagi and Thomas Peyrin, editors, *Advances in Cryptology – ASIACRYPT 2017*, volume 10211 of *LNCS*, pages 65–102. Springer, 2017. https://eprint.iacr.org/2017/815. 26, 27, 28

[7] Martin R. Albrecht, Rachel Player, and Sam Scott. On the concrete hardness of learning with errors. *J. Mathematical Cryptology*, 9(3):169–203, 2015. https://eprint.iacr.org/2015/046. 25, 26, 27

[8] Michael Alekhnovich. More on average case vs approximation complexity. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*, pages 298–307, 2003. 8

[9] Erdem Alkim, Yusuf Alper Bilgin, Murat Cenk, and François Gérard. Cortex-M4 optimizations for {R,M}LWE schemes. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(3):336–357, 2020. https://tches.iacr.org/index.php/TCHES/article/view/8593. 18

[10] Erdem Alkim, Léo Ducas, Thomas Pöppelmann, and Peter Schwabe. Post-quantum key exchange – a new hope. In *Proceedings of the 25th USENIX Security Symposium*, pages 327–343. USENIX Association, 2016. http://cryptojedi.org/papers/#newhope. 8, 12, 13, 15, 24, 25, 26, 27, 28, 34

[11] Yoshinori Aono, Yuntao Wang, Takuya Hayashi, and Tsuyoshi Takagi. Improved progressive BKZ algorithms and their precise cost estimation by sharp simulator. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology – EUROCRYPT 2016*, volume 9665 of *LNCS*, pages 789–819. Springer, 2016. https://eprint.iacr.org/2016/146. 25, 28

[12] Benny Applebaum, David Cash, Chris Peikert, and Amit Sahai. Fast cryptographic primitives and circular-secure encryption based on hard learning problems. In *Advances in Cryptology - CRYPTO 2009, 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009. Proceedings*, pages 595–618, 2009. 8

[13] Sanjeev Arora and Rong Ge. New algorithms for learning in presence of errors. In Luca Aceto, Monika Henzingeri, and Jiří Sgall, editors, *Automata, Languages and Programming*, volume 6755 of *LNCS*, pages 403–415. Springer, 2011. https://www.cs.duke.edu/~rongge/LPSN.pdf. 25

[14] Wojciech Banaszczyk. New bounds in some transference theorems in the geometry of numbers. *Mathematische Annalen*, 296(1):625–635, 1993. 32

[15] Abhishek Banerjee, Chris Peikert, and Alon Rosen. Pseudorandom functions and lattices. In David Pointcheval and Thomas Johansson, editors, *Advances in Cryptology – EUROCRYPT 2012*, volume 7237 of *LNCS*, pages 719–737. Springer, 2012. http://www.iacr.org/archive/eurocrypt2012/ 72370713/72370713.pdf. 8, 35

[16] Paul Barrett. Implementing the Rivest Shamir and Adleman public key encryption algorithm on a standard digital signal processor. In Andrew M. Odlyzko, editor, *Advances in Cryptology – CRYPTO '86*, volume 263 of

Lecture Notes in Computer Science, pages 311–323. Springer-Verlag Berlin Heidelberg, 1987. https://link.springer.com/chapter/10.1007/3-540-47721-7_24. 23

[17] Anja Becker, Léo Ducas, Nicolas Gama, and Thijs Laarhoven. New directions in nearest neighbor searching with applications to lattice sieving. In *SODA '16 Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete Algorithms*, pages 10–24. SIAM, 2016. https://eprint.iacr.org/2015/1128. 26, 27, 29

[18] Daniel J. Bernstein, Tanja Lange, and Peter Schwabe. The security impact of a new cryptographic library. In Alejandro Hevia and Gregory Neven, editors, *Progress in Cryptology – LATINCRYPT 2012*, volume 7533 of *LNCS*, pages 159–176. Springer, 2012. http://cryptojedi.org/papers/#coolnacl. 14

[19] Daniel J. Bernstein, Peter Schwabe, and Gilles Van Assche. Tweetable FIPS 202, 2015. https://keccak.team/2015/tweetfips202.html (accessed 2017-11-29). 14

[20] Guido Bertoni, Joan Daemen, Michaël Peeters, and Gilles Van Assche. The Keccak reference. Submission to the NIST SHA-3 competition, 2011. https://keccak.team/files/Keccak-reference-3.0.pdf. 14, 31

[21] Jean-François Biasse and Fang Song. Efficient quantum algorithms for computing class groups and solving the principal ideal problem in arbitrary degree number fields. In *SODA '16 Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete Algorithms*, pages 893–902. SIAM, 2016. http://fangsong.info/files/pubs/BS_SODA16.pdf. 31, 35

[22] Nina Bindel and John M. Schanck. Decryption failure is more likely after success. In Jintai Ding and Jean-Pierre Tillich, editors, *Post-Quantum Cryptography*, volume 12100 of *LNCS*, pages 206–225. Springer, 2020. https://eprint.iacr.org/2019/1392.pdf. 32

[23] Joppe Bos, Craig Costello, Léo Ducas, Ilya Mironov, Michael Naehrig, Valeria Nikolaenko, Ananth Raghunathan, and Douglas Stebila. Frodo: Take off the ring! practical, quantum-secure key exchange from LWE. In *CCS '16 Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1006–1018. ACM, 2016. https://eprint.iacr.org/2016/659. 12, 34

[24] Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Peter Schwabe, and Damien Stehlé. CRYSTALS – Kyber: a CCA-secure module-lattice-based KEM. In *2018 IEEE European Symposium on Security and Privacy, EuroS&P 2018*. IEEE, 2018. To appear. https://eprint.iacr.org/2017/634. 4, 11, 24

[25] Joppe W. Bos, Craig Costello, Michael Naehrig, and Douglas Stebila. Post-quantum key exchange for the TLS protocol from the ring learning with errors problem. In *2015 IEEE Symposium on Security and Privacy*, pages 553–570, 2015. https://eprint.iacr.org/2014/599. 12, 13

[26] Leon Botros, Matthias Kannwischer, and Peter Schwabe. Memory-efficient high-speed implementation of Kyber on Cortex-M4, 2019. http://cryptojedi.org/papers/#nttm4. 18

[27] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (Leveled) fully homomorphic encryption without bootstrapping. In *ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 309–325. ACM, 2012. https://eprint.iacr.org/2011/277. 19

[28] Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In *STOC '13 Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 575–584. ACM, 2013. http://arxiv.org/pdf/1306.0281. 13

[29] Leon Groot Bruinderink, Andreas Hülsing, Tanja Lange, and Yuval Yarom. Flush, Gauss, and Reload – a cache attack on the BLISS lattice-based signature scheme. In Benedikt Gierlichs and Axel Poschmann, editors, *Cryptographic Hardware and Embedded Systems – CHES 2016*, volume 9813 of *LNCS*, pages 323–345. Springer, 2016. https://eprint.iacr.org/2016/300. 13

[30] Peter Campbell, Michael Groves, and Dan Shepherd. Soliloquy: A cautionary tale. In *ETSI 2nd Quantum-Safe Crypto Workshop*, pages 1–9, 2014. https://docbox.etsi.org/workshop/2014/201410_CRYPTO/S07_Systems_and_Attacks/S07_Groves_Annex.pdf. 31

[31] Yuanmi Chen. *Lattice reduction and concrete security of fully homomorphic encryption*. PhD thesis, l'Université Paris Diderot, 2013. http://www.di.ens.fr/~ychen/research/these.pdf. 25, 26

[32] Yuanmi Chen and Phong Q. Nguyen. BKZ 2.0: Better lattice security estimates. In Dong Hoon Lee and Xiaoyun Wang, editors, *Advances in Cryptology – ASIACRYPT 2011*, volume 7073 of *LNCS*, pages 1–20. Springer, 2011. http://www.iacr.org/archive/asiacrypt2011/70730001/70730001.pdf. 25, 27

[33] Craig Costello, David Jao, Patrick Longa, Michael Naehrig, Joost Renes, and David Urbanik. Efficient compression of SIDH public keys. In Jean-Sébastien Coron and Jesper Buus Nielsen, editors, *Advances in Cryptology – EUROCRYPT 2017*, volume 10210 of *LNCS*, pages 679–706. Springer, 2017. https://eprint.iacr.org/2016/963. 33

[34] Ronald Cramer, Léo Ducas, Chris Peikert, and Oded Regev. Recovering short generators of principal ideals in cyclotomic rings. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology – EUROCRYPT 2016*, volume 9666 of *LNCS*, pages 559–585. Springer, 2016. https://eprint.iacr. org/2015/313. 31, 35

[35] Ronald Cramer, Léo Ducas, and Benjamin Wesolowski. Short Stickelberger class relations and application to Ideal-SVP. In Jean-Sébastien Coron and Jesper Buus Nielsen, editors, *Advances in Cryptology – EUROCRYPT 2017*, volume 10210 of *LNCS*, pages 324–348. Springer, 2017. https: //eprint.iacr.org/2016/885. 31, 35

[36] Dana Dachman-Soled, Léo Ducas, Huijing Gong, and Mélissa Rossi. Lwe with side information: Attacks and concrete security estimation. In Daniele Micciancio and Thomas Ristenpart, editors, *Advances in Cryptology – CRYPTO 2020*, volume 12171 of *LNCS*, pages 329–358. Springer, 2020. https://eprint.iacr.org/2020/292.pdf. 21, 27, 28, 30

[37] Jan-Pieter D'Anvers, Frederik Vercauteren, and Ingrid Verbauwhede. On the impact of decryption failures on the security of LWE/LWR based schemes. *IACR Cryptol. ePrint Arch.*, 2018:1089, 2018. 32

[38] The FPLLL development team. fplll, a lattice reduction library. Available at https://github.com/ fplll/fplll, 2017. 25

[39] Jintai Ding, Xiang Xie, and Xiaodong Lin. A simple provably secure key exchange scheme based on the learning with errors problem. IACR Cryptology ePrint Archive report 2012/688, 2012. https: //eprint.iacr.org/2012/688. 34

[40] Léo Ducas. Shortest vector from lattice sieving: a few dimensions for free. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 125–145. Springer, 2018. 21, 25, 26, 27, 28, 29

[41] Léo Ducas, Marc Stevens, and Wessel van Woerden. Advanced lattice sieving on GPUs, with tensor cores. Cryptology ePrint Archive, Report 2021/141, 2021. https://eprint.iacr.org/2021/141. 30

[42] Jan-Pieter D'Anvers, Qian Guo, Thomas Johansson, Alexander Nilsson, Frederik Vercauteren, and Ingrid Verbauwhede. Decryption failure attacks on ind-cca secure lattice-based schemes. In *IACR International Workshop on Public Key Cryptography*, pages 565–598. Springer, 2019. 32

[43] Jan-Pieter D'Anvers, Mélissa Rossi, and Fernando Virdia. (one) failure is not an option: Bootstrapping the search for failures in lattice-based encryption schemes. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 3–33. Springer, 2020. 33

[44] Kirsten Eisenträger, Sean Hallgren, Alexei Kitaev, and Fang Song. A quantum algorithm for computing the unit group of an arbitrary degree number field. In *STOC '14 Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 293–302. ACM, 2014. http://www.personal.psu. edu/kxe8/unitgroup.pdf. 31

[45] Thomas Espitau, Pierre-Alain Fouque, Benoït Gérard, and Mehdi Tibouchi. Side-channel attacks on BLISS lattice-based signatures: Exploiting branch tracing against strongswan and electromagnetic emanations in microcontrollers. In *CCS '17 Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1857–1874. ACM, 2017. https://eprint.iacr.org/ 2017/505. 13

[46] Scott Fluhrer. Cryptanalysis of ring-LWE based key exchange with key share reuse. IACR Cryptology ePrint Archive report 2016/085, 2016. https://eprint.iacr.org/2016/085. 33

[47] Eiichiro Fujisaki and Tatsuaki Okamoto. Secure integration of asymmetric and symmetric encryption schemes. In *Advances in Cryptology - CRYPTO '99*, pages 537–554, 1999. https://link.springer. com/chapter/10.1007/3-540-48405-1_34. 4, 10, 20

[48] Nicolas Gama and Phong Nguyen. Predicting lattice reduction. In Nigel Smart, editor, *Advances in Cryptology – EUROCRYPT 2008*, volume 4965 of *LNCS*, pages 31–51. Springer, 2008. https://www.iacr.org/archive/eurocrypt2008/49650031/49650031.pdf. 26

[49] Nicolas Gama, Phong Q Nguyen, and Oded Regev. Lattice enumeration using extreme pruning. In Henri Gilbert, editor, *Advances in Cryptology – EUROCRYPT 2010*, volume 6110 of *LNCS*, pages 257–278. Springer, 2010. http://www.iacr.org/archive/eurocrypt2010/66320257/66320257.pdf. 25

[50] Matthew Gretton-Dann. Introducing 2017's extensions to the Arm architecture, 2017. https://community.arm.com/processors/b/blog/posts/introducing-2017s-extensionsto-the-arm-architecture. 15

[51] Tim Güneysu, Tobias Oder, Thomas Pöppelmann, and Peter Schwabe. Software speed records for lattice-based signatures. In Philippe Gaborit, editor, *Post-Quantum Cryptography*, volume 7932 of *LNCS*, pages 67–82. Springer, 2013. http://cryptojedi.org/papers/#lattisigns. 15

[52] J. Alex Halderman, Seth D. Schoen, Nadia Heninger, William Clarkson, William Paul, Joseph A. Calandrino, Ariel J. Feldman, Jacob Appelbaum, and Edward W. Felten. Lest we remember: Cold boot attacks on encryption keys. In *Proceedings of the 17th USENIX Security Symposium*, pages 45–60. USENIX Association, 2008. https://www.usenix.org/legacy/event/sec08/tech/full_papers/ halderman/halderman.pdf. 16

[53] Guillaume Hanrot, Xavier Pujol, and Damien Stehlé. Terminating BKZ. IACR Cryptology ePrint Archive report 2011/198, 2011. https://eprint.iacr.org/2011/198. 25

[54] Jeffrey Hoffstein, Jull Pipher, and Joseph H. Silverman. NTRU: a ring-based public key cryptosystem. In Joe P. Buhler, editor, *Algorithmic number theory*, volume 1423 of *LNCS*, pages 267–288. Springer, 1998. https://www.securityinnovation.com/uploads/Crypto/ANTS97.ps.gz. 8, 34

[55] Dennis Hofheinz, Kathrin Hövelmanns, and Eike Kiltz. A modular analysis of the Fujisaki-Okamoto transformation. In Yael Kalai and Leonid Reyzin, editors, *Theory of Cryptography*, LNCS, pages 341–371. Springer, 2017. https://eprint.iacr.org/2017/604. 14, 20

[56] Nick Howgrave-Graham. A hybrid lattice-reduction and meet-in-the-middle attack against NTRU. In Alfred Menezes, editor, *Advances in Cryptology – CRYPTO 2007*, volume 4622 of *LNCS*, pages 150–169. Springer, 2007. http://www.iacr.org/archive/crypto2007/46220150/46220150.pdf. 13

[57] David Jao and Luca De Feo. Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies. In Bo-Yin Yang, editor, *Post-Quantum Cryptography – PQCrypto 2011*, volume 7071 of *LNCS*, pages 19–34. Springer, 2011. https://eprint.iacr.org/2011/506. 33

[58] Matthias Kannwischer, Joost Rijneveld, Peter Schwabe, Douglas Stebila, and Thom Wiggers. Pqclean: clean, portable, tested implementations of post-quantum cryptography. 16

[59] Matthias J. Kannwischer, Joost Rijneveld, Peter Schwabe, and Ko Stoffelen. pqm4: Testing and benchmarking NIST PQC on ARM Cortex-M4. Workshop Record of the Second PQC Standardization Conference, 2019. https://cryptojedi.org/papers/#pqm4. 2, 18

[60] Paul Kirchner and Pierre-Alain Fouque. An improved BKW algorithm for LWE with applications to cryptography and lattices. In Rosario Gennaro and Matthew Robshaw, editors, *Advances in Cryptology – CRYPTO 2015*, volume 9215 of *LNCS*, pages 43–62. Springer, 2015. http://www.iacr.org/archive/crypto2015/92160264/92160264.pdf. 25

[61] Paul Kirchner and Pierre-Alain Fouque. Revisiting lattice attacks on overstretched NTRU parameters. In Jean-Sébastien Coron and Jesper Buus Nielsen, editors, *Advances in Cryptology – EUROCRYPT 2017*, volume 10210 of *LNCS*, pages 3–26. Springer, 2017. https://www.di.ens.fr/~fouque/ euro17a.pdf. 35

[62] Thijs Laarhoven. *Search problems in cryptography*. PhD thesis, Eindhoven University of Technology, 2015. http://www.thijs.com/docs/phd-final.pdf. 26

[63] Thijs Laarhoven. Sieving for shortest vectors in lattices using angular locality-sensitive hashing. In Rosiario Gennaro and Matthew Robshaw, editors, *Advances in Cryptology – CRYPTO 2015*, volume 9216 of *LNCS*, pages 3–22. Springer, 2015. http://www.iacr.org/archive/crypto2015/92160123/ 92160123.pdf. 26

[64] Thijs Laarhoven and Artur Mariano. Progressive lattice sieving. In *International Conference on PostQuantum Cryptography*, pages 292–311. Springer, 2018. 28

[65] Thijs Laarhoven, Michele Mosca, and Joop van de Pol. Finding shortest lattice vectors faster using quantum search. *Designs, Codes and Cryptography*, 77(2):375–400, 2015. https://eprint.iacr.org/ 2014/907. 26

[66] Adam Langley. Maybe skip SHA-3. Blog post on ImperialViolet, 2017. https://www.imperialviolet.org/2017/05/31/skipsha3.html. 15

[67] Adeline Langlois and Damien Stehlé. Worst-case to average-case reductions for module lattices. *Designs, Codes and Cryptography*, 75(3):565–599, 2015. https://eprint.iacr.org/2012/090. 4, 12, 19

[68] Mingjie Liu and Phong Q Nguyen. Solving bdd by enumeration: An update. In *Cryptographers' Track at the RSA Conference*, pages 293–309. Springer, 2013. 30

[69] Vadim Lyubashevsky. Standardizing lattice crypto and beyond. Slides of the talk given by Vadim Lyubashevsky at PQCrypto 2017, 2017. https://2017.pqcrypto.org/conference/slides/pqc_ 2017_lattice.pdf. 34

[70] Vadim Lyubashevsky, Daniele Micciancio, Chris Peikert, and Alon Rosen. SWIFFT: A modest proposal for FFT hashing. In Kaisa Nyberg, editor, *Fast Software Encryption – FSE 2008,*, volume 5086 of *LNCS*, pages 54–72. Springer, 2008. https://www.eecs.harvard.edu/~alon/PAPERS/lattices/ swifft.pdf. 12, 35

[71] Vadim Lyubashevsky, Adriana Palacio, and Gil Segev. Public-key cryptographic primitives provably as secure as subset sum. In *Theory of Cryptography, 7th Theory of Cryptography Conference, TCC 2010, Zurich, Switzerland, February 9-11, 2010. Proceedings*, pages 382–400, 2010. 8

[72] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. In Henri Gilbert, editor, *Advances in Cryptology – EUROCRYPT 2010*, volume 6110 of *LNCS*, pages 1–23. Springer, 2010. http://www.iacr.org/archive/eurocrypt2010/66320288/66320288. pdf. 7, 12, 34, 35

[73] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. Slides of the talk given by Chris Peikert at Eurocrypt 2010, 2010. http://crypto.rd.francetelecom.com/events/eurocrypt2010/talks/slides-ideal-lwe.pdf. 7

[74] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. *Journal of the ACM*, 60(6):43:1–43:35, 2013. http://www.cims.nyu.edu/~regev/papers/ ideal-lwe.pdf. 7

[75] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. A toolkit for Ring-LWE cryptography. In Thomas Johansson and Phong Q. Nguyen, editors, *Advances in Cryptology – EUROCRYPT 2013*, volume 7881 of *LNCS*, pages 35–54. Springer, 2013. http://www.iacr.org/archive/eurocrypt2013/ 78810035/78810035.pdf. 34

[76] Vadim Lyubashevsky and Gregor Seiler. NTTRU: Truly fast NTRU using NTT. *TCHES*, 2019. https://eprint.iacr.org/2019/040. 34

[77] Artur Mariano, Thijs Laarhoven, and Christian Bischof. A parallel variant of LDSieve for the SVP on lattices. In *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 23–30. IEEE, 2017. 26, 27

[78] Daniele Micciancio and Panagiotis Voulgaris. Faster exponential time algorithms for the shortest vector problem. In *SODA '10 Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1468–1480. SIAM, 2010. https://cseweb.ucsd.edu/~daniele/papers/Sieve. pdf. 26

[79] Peter L. Montgomery. Modular multiplication without trial division. *Mathematics of Computation*, 44(170):519–521, 1985. http://www.ams.org/journals/mcom/1985-44-170/S0025-5718-1985-0777282-X/S0025-5718-1985-0777282-X.pdf. 23

[80] Tamalika Mukherjee and Noah Stephens-Davidowitz. Lattice reduction for modules, or how to reduce modulesvp to modulesvp. In *Annual International Cryptology Conference*, pages 213–242. Springer, 2020. 30

[81] National Institute of Standards and Technology. FIPS PUB 202 – SHA-3 standard: Permutationbased hash and extendable-output functions, 2015. http://nvlpubs.nist.gov/nistpubs/FIPS/ NIST.FIPS.202.pdf. 11, 14

[82] Phong Q. Nguyen and Thomas Vidick. Sieve algorithms for the shortest vector problem are practical. *Journal of Mathematical Cryptology*, 2(2):181–207, 2008. ftp://ftp.di.ens.fr/pub/users/ pnguyen/JoMC08.pdf. 26

[83] Tobias Oder, Tobias Schneider, Thomas Pöppelmann, and Tim Güneysu. Practical CCA2-secure and masked Ring-LWE implementation. IACR Cryptology ePrint Archive report 2016/1109, 2016. https://eprint.iacr.org/2016/1109. 24

[84] Chris Peikert. Public-key cryptosystems from the worst-case shortest vector problem, 2009. https://web.eecs.umich.edu/~cpeikert/pubs/svpcrypto.pdf (full version of [85]). 8, 12, 42

[85] Chris Peikert. Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In *STOC '09 Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 333–342. ACM, 2009. See also full version [84]. 42

[86] Chris Peikert. Lattice cryptography for the Internet. In Michele Mosca, editor, *Post-Quantum Cryptography*, volume 8772 of *LNCS*, pages 197–219. Springer, 2014. http://web.eecs.umich.edu/ ~cpeikert/pubs/suite.pdf. 34

[87] Chris Peikert. How (not) to instantiate Ring-LWE. In Vassilis Zikas and Roberto De Prisco, editors, *Security and Cryptography for Networks*, volume 9841 of *LNCS*, pages 411–430. Springer, 2016. https://web.eecs.umich.edu/~cpeikert/pubs/instantiate-rlwe.pdf. 35

[88] Chris Peikert, Vinod Vaikuntanathan, and Brent Waters. A framework for efficient and composable oblivious transfer. In David A. Wagner, editor, *Advances in Cryptology – CRYPTO 2008*, volume 5157 of *LNCS*, pages 554–571. Springer, 2008. https://www.iacr.org/archive/crypto2008/51570556/51570556.pdf. 34

[89] Peter Pessl, Leon Groot Bruinderink, and Yuval Yarom. To BLISS-B or not to be – attacking strongSwan's implementation of post-quantum signatures. In *CCS '17 Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1843–1855. ACM, 2017. https://eprint.iacr.org/2017/490. 13

[90] Thomas Pöppelmann and Tim Güneysu. Towards practical lattice-based public-key encryption on reconfigurable hardware. In Tanja Lange, Kristin Lauter, and Petr Lisoněk, editors, *Selected Areas in Cryptography – SAC 2013*, volume 8282 of *LNCS*, pages 68–85. Springer, 2013. https://www.ei.rub.de/media/sh/veroeffentlichungen/2013/08/14/lwe_encrypt.pdf. 8, 12

[91] Thomas Pornin. BearSSL – a smaller SSL/TLS library, 2018. https://bearssl.org/ (accessed 201903-15). 23

[92] Robert Primas, Peter Pessl, and Stefan Mangard. Single-trace side-channel attacks on masked latticebased encryption. In Wieland Fischer and Naofumi Homma, editors, *Cryptographic Hardware and Embedded Systems – CHES 2017*, volume 10529 of *LNCS*, pages 513–533. Springer, 2017. https://eprint.iacr.org/2017/594. 24

[93] Prasanna Ravi, Debapriya Basu Roy, Shivam Bhasin, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Number "not used" once - practical fault attack on pqm4 implementations of nist candidates. IACR Cryptology ePrint Archive report 2018/211, 2018. https://eprint.iacr.org/2018/211. 36

[94] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *STOC '05 Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 84–93. ACM, 2005. Preliminary version of [95]. 8, 13, 34

[95] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM*, 56(6):34, 2009. http://www.cims.nyu.edu/~regev/papers/qcrypto.pdf. 8, 43

[96] Sujoy Sinha Roy, Frederik Vercauteren, Nele Mentens, Donald Donglong Chen, and Ingrid Verbauwhede. Compact Ring-LWE cryptoprocessor. In Lejla Batina and Matthew Robshaw, editors, *Cryptographic Hardware and Embedded Systems – CHES 2014*, volume 8731 of *LNCS*, pages 371–391. Springer, 2014. http://www.iacr.org/archive/ches2014/87310183/87310183.pdf. 12

[97] Tsunekazu Saito, Keita Xagawa, and Takashi Yamakawa. Tightly-secure key-encapsulation mechanism in the quantum random oracle model. IACR Cryptology ePrint Archive report 2017/1005, 2017. https://eprint.iacr.org/2017/1005. 20

[98] Claus-Peter Schnorr and Martin Euchner. Lattice basis reduction: improved practical algorithms and solving subset sum problems. *Mathematical programming*, 66(1-3):181–199, 1994. http://www.csie.nuk.edu.tw/~cychen/Lattices/Lattice%20Basis%20Reduction_%20Improved%20Practical%20Algorithms%20and%20Solving%20Subset%20Sum%20Problems.pdf. 25

[99] Yang Yu and Léo Ducas. Second order statistical behavior of lll and bkz. In *International Conference on Selected Areas in Cryptography*, pages 3–22. Springer, 2017. 28, 30

[100] Shuai Zhou, Haiyang Xue, Daode Zhang, Kunpeng Wang, Xianhui Lu, Bao Li, and Jingnan He. Preprocess-then-ntt technique and its applications to KYBER and NEWHOPE. IACR Cryptology ePrint Archive report 2018/995, 2018. https://eprint.iacr.org/2018/995. 36