Speeding up the Number Theoretic Transform for Faster Ideal Lattice-Based Cryptography

Patrick Longa(✉) and Michael Naehrig

Microsoft Research, Redmond, USA
{plonga,mnaehrig}@microsoft.com

Abstract. The Number Theoretic Transform (NTT) provides efficient algorithms for cyclic and nega-cyclic convolutions, which have many applications in computer arithmetic, e.g., for multiplying large integers and large degree polynomials. It is commonly used in cryptographic schemes that are based on the hardness of the Ring Learning With Errors (R-LWE) problem to efficiently implement modular polynomial multiplication.

We present a new modular reduction technique that is tailored for the special moduli required by the NTT. Based on this reduction, we speed up the NTT and propose faster, multi-purpose algorithms. We present two implementations of these algorithms: a portable C implementation and a high-speed implementation using assembly with AVX2 instructions. To demonstrate the improved efficiency in an application example, we benchmarked the algorithms in the context of the R-LWE key exchange protocol that has recently been proposed by Alkim, Ducas, Pöppelmann and Schwabe. In this case, our C and assembly implementations compute the full key exchange 1.44 and 1.21 times faster, respectively. These results are achieved with full protection against timing attacks.

Keywords: Post-quantum cryptography · Number Theoretic Transform (NTT) · Ring Learning With Errors (R-LWE) · Fast modular reduction · Efficient implementation

1 Introduction

Fast Fourier Transform (FFT) algorithms to compute the Discrete Fourier Transform (DFT) have countless applications ranging from digital signal processing to the fast multiplication of large integers. The cyclic convolution of two integer sequences of length n can be computed by applying an FFT algorithm to both, then multiplying the resulting DFT sequences of length n coefficient-wise and transforming the result back via an inverse FFT. This operation corresponds to the product of the corresponding polynomials modulo $X^n - 1$, and for large n, a computation via FFTs as above was suggested to be used in the ring-based encryption scheme NTRUEncrypt in [15].

S. Foresti and G. Persiano (Eds.): CANS 2016, LNCS 10052, pp. 124–139, 2016.
DOI: 10.1007/978-3-319-48965-0_8

When the sequence (or polynomial) coefficients are specialized to come from a finite field, the DFT is called the Number Theoretic Transform (NTT) [8] and can be computed with FFT algorithms that work over this specific finite field. Polynomial multiplication over a finite field is one of the fundamental operations required in cryptographic schemes based on the Ring Learning With Errors (R-LWE) problem, and the NTT has shown to be a powerful tool that enables this operation to be computed in quasi-polynomial complexity.

R-LWE-Based Cryptography. Since its introduction by Regev [28], the Learning With Errors (LWE) problem has been used as the foundation for many new lattice-based constructions with a variety of cryptographic functionalities. It is currently believed to be sufficiently hard, even for attackers running a large scale quantum computer. Hence cryptographic schemes with security based on the hardness of the LWE problem are promising candidates for post-quantum (or quantum-safe) cryptography.

The Ring LWE (R-LWE) problem, introduced by Lyubashevsky, Peikert and Regev [20], is a special instance of the LWE problem that is essentially obtained by adding a ring structure to the underlying lattice. R-LWE-based schemes have been proposed for public-key encryption [20,24,31], digital signatures [11,19], and key exchange [2,5,10,24,32]. Furthermore, the most efficient proposals for (fully) homomorphic encryption are also based on R-LWE, e.g., [6].

The advantage of R-LWE over LWE is a significant increase in efficiency. When working with vectors of dimension n, it allows a factor n space reduction and the possibility of using FFT algorithms to compute polynomial products instead of matrix-vector or matrix-matrix operations; this leads to an improvement from roughly n^2 base ring multiplications to roughly $n \log n$ such multiplications.

One particularly efficient parameter instantiation in the context of R-LWE is such that the dimension n is a power of 2 and polynomial products are taken modulo the $2n$-th cyclotomic polynomial $X^n + 1$ with coefficients modulo a prime q. Here, the polynomial product corresponds to a nega-cyclic convolution of the coefficient sequences. In this setting, the NTT is usually computed with a special type of FFT algorithm that can be used efficiently when q is a prime that satisfies the congruence condition $q \equiv 1 \bmod 2n$ (cf. [21, Sect. 2.1]), which in turn means that the underlying finite field contains primitive $2n$-th roots of unity. Many state-of-the-art instantiations of R-LWE-based cryptography choose n and q as above in order to harness the efficiency of the NTT; for example, the BLISS signature implementations (I-IV) set $n = 512$ and $q = 12289$ [11] and the fastest R-LWE-based key exchange implementation to date sets $n = 1024$ and $q = 12289$ [2].

Our Contributions. We present a new modular reduction algorithm for the special moduli that are required to invoke the NTT. While this new routine can be used to replace existing modular reduction algorithms and give standalone performance improvements, we further show that calling it inside a modified

NTT algorithm can give rise to additional speedups. We illustrate these improvements by providing and benchmarking both our portable C and AVX2 assembly implementations (see Sect. 5 for complete details). Our software is publicly available as part of the LatticeCrypto library [18].

Given the ubiquity of the NTT in (both the existing and foreseeable) high-speed instantiations of R-LWE-based primitives, we emphasize that an improved NTT simultaneously improves a large portion of all lattice-based cryptographic proposals. While our algorithm will give a solid speedup to signature schemes like Lyubashevsky's [19] and BLISS [11], it will give a more drastic overall improvement in common encryption and key exchange schemes. In these scenarios, there are different ways of removing the need for obtaining high-precision samples from a Gaussian distribution [22], for example, the number of R-LWE samples per secret can be bounded, or one can use the Kullback-Leibler or Renyi divergences [3]. Subsequently, the cost of sampling the error distribution decreases dramatically, and the NTT becomes the bottleneck of the overall computation.

To highlight the practical benefits of the new approach in an example of a cryptographic protocol, we implemented the recent key exchange instantiation due to Alkim, Ducas, Pöppelmann and Schwabe [2], and show that the overall key exchange is approximately 1.44 times faster (portable C implementation) and 1.21 times faster (AVX2 assembly implementation) using our improved NTT.

Beyond the faster modular reduction itself, the specific improvements over the approach in [2] that have led to this speedup are as follows:

- The new modular reduction algorithm allows coefficients to grow up to 32 bits in size, which eliminates the need for modular reductions after any addition during the NTT. As a consequence, reductions are only carried out after multiplications.
- The new modular reduction is very flexible and enables efficient implementations using either integer arithmetic or floating point arithmetic. Since it minimizes the use of multiplications, using the higher throughput of floating point instructions on the latest Intel processors does not have as big an impact as for more multiplication-heavy methods like Montgomery reduction. Hence, the method is especially attractive for implementations with a focus on simplicity, particularly in plain C.
- Related to the previous point, our implementation uses signed integer arithmetic in the NTT. This allows for signed integers to represent error polynomials and secret keys, which saves conversions from negative to positive integers (e.g., this reduces the number of additions during error sampling and before modular reductions in the NTT).
- We show how to merge the scaling by n^{-1} with our conversion from redundant to standard integer representation at the end of the inverse NTT. In addition, by pulling this conversion into the last stage of the inverse NTT, we eliminate $n/2$ multiplications and reductions, all at the cost of precomputing only two integers.

Organization. Section 2 gives the background on R-LWE and the NTT. Section 3 contains our two main contributions: the improved modular reduction and NTT algorithms. Section 4 revises the details in the R-LWE key exchange scheme from [2], which is used as a case study to give a practical instance where our improved NTT gives rise to faster cryptography. Finally, Sect. 5 provides a performance analysis and benchmarks.

2 Preliminaries

This section provides details about the ring structure in the R-LWE setting, the NTT, and the FFT algorithm to compute the NTT and its inverse. The original proposal of R-LWE [20] restricts to cyclotomic rings, i.e. rings generated over the integers by primitive roots of unity. We immediately focus on 2-power cyclotomic rings as this is the most commonly used case and seems to provide the most efficient arithmetic.

2.1 The Ring Learning with Errors (R-LWE) Setting

Let $N = 2^d$, $d > 1$ be a power of two and let $n = \varphi(N) = 2^{d-1} = N/2$. Then the N-th cyclotomic polynomial is given by $\Phi_N(X) = X^n + 1$. Let R be the ring of cyclotomic integers, i.e. $R = \mathbb{Z}[X]/(\Phi_N(X)) = \mathbb{Z}[X]/(X^n + 1)$. Any element $a \in R$ can be written as $a = \sum_{i=0}^{n-1} a_i X^i$, $a_i \in \mathbb{Z}$. Furthermore, let $q \in \mathbb{Z}$ be a positive integer modulus such that $q \equiv 1 \pmod N$. The quotient ring $R/(q)$ is isomorphic to $R_q = \mathbb{Z}_q[X]/(X^n + 1)$ and for any $a \in R_q$, we write $a = \sum_{i=0}^{n-1} a_i X^i$, $a_i \in \mathbb{Z}_q$. We use the same symbol a to also denote both the coefficient vector $a = (a_0, a_1, \ldots, a_{n-1}) \in \mathbb{Z}_q^n$ and the sequence $a = (a[0], a[1], \ldots, a[n-1]) \in \mathbb{Z}_q^n$.

2.2 The Number Theoretic Transform (NTT)

The NTT is a specialized version of the discrete Fourier transform, in which the coefficient ring is taken to be a finite field (or ring) containing the right roots of unity. It can be viewed as an exact version of the complex DFT, avoiding round-off errors for exact convolutions of integer sequences. While Gauss apparently used similar techniques already in [12], laying the ground work for modern FFT algorithms to compute the DFT and therefore the NTT is usually attributed to Cooley and Tukey's seminal paper [8].

Notation and Background. With parameters as above, i.e. n being a power of 2 and q a prime with $q \equiv 1 \pmod{2n}$, let $a = (a[0], ..., a[n-1]) \in \mathbb{Z}_q^n$, and let ω be a primitive n-th root of unity in $\mathbb{Z}_q$, which means that $\omega^n \equiv 1 \pmod q$. The forward transformation $\tilde{a} = \text{NTT}(a)$ is defined as $\tilde{a}[i] = \sum_{j=0}^{n-1} a[j]\omega^{ij} \bmod q$ for $i = 0, 1, ..., n-1$. The inverse transformation is given by $b = \text{INTT}(\tilde{a})$, where $b[i] = n^{-1}\sum_{j=0}^{n-1}\tilde{a}[j]\omega^{-ij} \bmod q$ for $i = 0, 1, ..., n-1$, and we have $\mathrm{INTT}(\mathrm{NTT}(a)) = a$.

As mentioned above, the NTT can be used directly to perform the main operation in R-LWE-based cryptography, that is, polynomial multiplication in $R_q = \mathbb{Z}_q[X]/(X^n + 1)$. However, since applying the NTT transform as described above provides a cyclic convolution, computing $c = a \cdot b \bmod (X^n + 1)$ with two polynomials a and b would require applying the NTT of length $2n$ and thus n zeros to be appended to each input; this effectively doubles the length of the inputs and also requires the computation of an explicit reduction modulo $X^n + 1$. To avoid these issues, one can exploit the *negative wrapped convolution* [21]: let ψ be a primitive $2n$-th root of unity in $\mathbb{Z}_q$ such that $\psi^2 = \omega$, and let $a = (a[0], ..., a[n-1])$, $b = (b[0], ..., b[n-1]) \in \mathbb{Z}_q^n$ be two vectors. Also, define $\hat{a} = (a[0], \psi a[1]..., \psi^{n-1}a[n-1])$ and $\hat{b} = (b[0], \psi b[1]..., \psi^{n-1}b[n-1])$. The negative wrapped convolution of a and b is defined as $c = (1, \psi^{-1}, \psi^{-2}, ..., \psi^{-(n-1)}) \circ \mathrm{INTT}(\mathrm{NTT}(\hat{a}) \circ \mathrm{NTT}(\hat{b}))$, where $\circ$ denotes component-wise multiplication. This operation satisfies $c = a \cdot b$ in R_q.

Previous Optimizations. Some additional optimizations are available to the NTT-based polynomial multiplication. Previous works explain how to merge multiplications by the powers of ω with the powers of ψ and ψ^{-1} inside the NTT. Consequently, important savings can be achieved by precomputing and storing in memory the values related to these parameters. In particular, Roy et al. [29] showed how to merge the powers of ψ with the powers of ω in the forward transformation. This merging did not pose any difficulty in the case of the well-known *decimation-in-time* NTT, which is based on the Cooley-Tukey butterfly [8] that was used in the first implementations of R-LWE-based schemes. Similarly, Pöppelmann et al. [26] showed how to merge the powers of ψ^{-1} with the powers of ω in the inverse transformation. In this case, however, it was necessary to switch from a decimation-in-time NTT to a *decimation-in-frequency* NTT [13], which is based on the Gentleman-Sande (GS) butterfly. In this work we exploit the combination of both transformations for optimal performance.

Other optimizations focus on the NTT's butterfly computation. Relevant examples are the use of precomputed quotients, as exploited in Shoup's butterfly algorithm [30], and the use of redundant representations that enable the elimination of several conditional modular corrections, as shown by Harvey [14]. In particular, Harvey showed how to apply the latter technique on Shoup's butterfly and on a butterfly variant based on Montgomery arithmetic. In Sect. 5, we compare our improved NTT algorithms with the approaches by Melchor et al. [1] and Alkim et al. [2], both of which adopted and specialized Harvey's butterfly algorithms.

Several works in the literature (e.g., [2,17,25,29]) have applied a relatively expensive reordering or bit-reversal step before or after the NTT computation. This is due to the restrictive nature of certain forward and inverse algorithms that only accept inputs in standard ordering and produce results in bit-reversed ordering. However, Chu and George [7] showed how to also derive forward and inverse FFT algorithms working for the reversed case, i.e., accepting inputs in bit-reversed ordering and producing outputs in standard ordering. Accordingly, [26] adapted and suitably combined the algorithms in the context of NTTs in order to eliminate the need of the bit-reversal step.

From hereon, we denote by $\texttt{NTT} := \mathrm{NTT}_{\mathrm{CT},\Psi_{\mathrm{rev}}}$ an algorithm that computes the forward transformation based on the Cooley-Tukey butterfly that absorbs the powers of ψ in bit-reversed ordering. This function receives the inputs in standard ordering and produces a result in bit-reversed ordering. Similarly, we denote by $\texttt{INTT} := \mathrm{INTT}_{\mathrm{GS},\Psi^{-1}_{\mathrm{rev}}}$ an algorithm computing the inverse transformation based on the Gentleman-Sande butterfly that absorbs the powers of ψ^{-1} in the bit-reversed ordering. This function receives the inputs in bit-reversed ordering and produces an output in standard ordering. Following Pöppelmann et al. [26], the combination of these two functions eliminates any need for a bit-reversal step. Optimized algorithms for the forward and inverse NTT are presented in Algorithms 1 and 2, respectively. These algorithms are based on the ones detailed in [26, Appendix A.1]. Note that we have applied a few modifications and corrected some typos.

Pöppelmann et al. [26] avoid the final scaling by n^{-1} during the inverse NTT by shifting the computation to a polynomial transformation that is (in their target application of BLISS signatures) assumedly performed offline. In general, however, that assumption does not necessarily hold; for example, in [2], all of the polynomials to be multiplied are generated *fresh* per key exchange connection. Accordingly, Algorithm 2 includes scaling by n^{-1}.

Algorithm 1. Function NTT based on the Cooley-Tukey (CT) butterfly.

Input: A vector $a = (a[0], a[1], ..., a[n-1]) \in \mathbb{Z}_q^n$ in standard ordering, where q is a prime such that $q \equiv 1 \bmod 2n$ and n is a power of two, and a precomputed table $\Psi_{rev} \in \mathbb{Z}_q^n$ storing powers of ψ in bit-reversed order.
Output: $a \leftarrow \texttt{NTT}(a)$ in bit-reversed ordering.

```
$t = n$
for ($m = 1$; $m < n$; $m = 2m$) do
    $t = t/2$
    for ($i = 0$; $i < m$; $i$++) do
        $j_1 = 2 \cdot i \cdot t$
        $j_2 = j_1 + t - 1$
        $S = \Psi_{rev}[m + i]$
        for ($j = j_1$; $j \leq j_2$; $j$++) do
            $U = a[j]$
            $V = a[j + t] \cdot S$
            $a[j] = U + V \bmod q$
            $a[j + t] = U - V \bmod q$
return $a$
```

Algorithm 2. Function INTT based on the Gentleman-Sande (GS) butterfly.

Input: A vector $a = (a[0], a[1], ..., a[n-1]) \in \mathbb{Z}_q^n$ in bit-reversed ordering, where q is a prime such that $q \equiv 1 \bmod 2n$ and n is a power of two, and a precomputed table $\Psi_{rev}^{-1} \in \mathbb{Z}_q^n$ storing powers of ψ^{-1} in bit-reversed order.
Output: $a \leftarrow \texttt{INTT}(a)$ in standard ordering.

```
$t = 1$
for ($m = n$; $m > 1$; $m = m/2$) do
    $j_1 = 0$
    $h = m/2$
    for ($i = 0$; $i < h$; $i{+}{+}$) do
        $j_2 = j_1 + t - 1$
        $S = \Psi_{rev}^{-1}[h + i]$
        for ($j = j_1$; $j \leq j_2$; $j{+}{+}$) do
            $U = a[j]$
            $V = a[j + t]$
            $a[j] = U + V \bmod q$
            $a[j + t] = (U - V) \cdot S \bmod q$
        $j_1 = j_1 + 2t$
    $t = 2t$
for ($j = 0$; $j < n$; $j{+}{+}$) do
    $a[j] = a[j] \cdot n^{-1} \bmod q$
return $a$
```

3 Modular Reduction and Speeding up the NTT

Most FFT algorithms to compute the NTT over a finite field or ring need certain roots of unity. In the specific setting discussed in the previous section, one needs primitive $2n$-th roots of unity to exist[1] modulo q, which imposes a congruence condition on q, namely $q \equiv 1 \pmod{2n}$. The parameters for R-LWE-based cryptosystems tend to have relatively large dimension n and relatively small moduli q, which means that moduli satisfying the congruence have the form $q = k \cdot 2^m + 1$, where $2n \mid 2^m$ and $k \geq 3$ is a very small integer.

Modular Reduction. In this section, we introduce a new modular reduction method for moduli of this special shape. We note that it works similarly for any modulus of the form $k \cdot 2^m \pm l$, where k and l are small positive integers such that $k \geq 3$ and $l \geq 1$. However, for ease of exposition and to focus on the case most relevant in the context of the NTT, we only treat the case $q = k \cdot 2^m + 1$. When k is odd and $2^m > k$, these numbers are known as Proth numbers [27], and a general algorithm for reduction modulo such integers is discussed in [9, Section 9.2.3].

Let $0 \leq a, b < q$ be two integers modulo q and let $C = a \cdot b$ be their integer product. Then $0 \leq C < q^2 = k^2 2^{2m} + k 2^{m+1} + 1$. The goal is to reduce C modulo

[1] For an algorithm that does not require such roots, but has the disadvantage of needing to pad the inputs to double length to compute nega-cyclic convolutions, see Nussbaumers algorithm ([23] and [16, Exercise 4.6.4.59]).

q using the special shape of q, namely using the fact that $k2^m \equiv -1 \pmod{q}$. Write $C = C_0 + 2^m C_1$, where $0 \le C_0 < 2^m$. Then $0 \le C_1 = (C - C_0)/2^m < k^2 2^m + 2k + 1/2^m = kq + k + 1/2^m$. We have that $kC \equiv kC_0 - C_1 \pmod{q}$, and given the above bounds for C_0 and C_1, it follows that the integer $kC_0 - C_1$ has absolute value bounded by $|kC_0 - C_1| < (k + 1/2^m)q$. As k is a small integer, the value $kC_0 - C_1$ can be brought into the range $[0, q)$ by adding or subtracting a small multiple of q. The maximal value for C is $(q-1)^2 = k^2 2^{2m}$, in which case $C_0 = 0$ and $C_1 = k^2 2^m = k(q-1)$, meaning that $(k-1)q$ must be added to $kC_0 - C_1$ to fully reduce the result. In our application to the NTT, however, we do not intend to perform this final reduction into $[0, q)$ throughout the computation, but rather only at the very end of the algorithm. We are therefore content with the output of the function `K-RED` defined as follows:

```
function K-RED($C$)
    $C_0 \leftarrow C \bmod 2^m$
    $C_1 \leftarrow C/2^m$
    return $kC_0 - C_1$
```

The function `K-RED` can take any integer C as input. It then returns an integer D such that $D \equiv kC \pmod{q}$ and $|D| < q + |C|/2^m$. Although this function alone does not properly reduce the value C modulo q, we still call it a reduction because it brings D close to the desired range; note that for $|C| > (2^m/(2^m - 1))q$, we have $|D| < |\texttt{K-RED}(C)|$, i.e. it reduces the size of C. As a specific example, take $q = 12289 = 3 \cdot 2^{12} + 1$. Then $k = 3$ and `K-RED` returns $3C_0 - C_1 \equiv 3C \pmod{q}$ using the equivalence $3 \cdot 2^{12} \equiv -1 \pmod{q}$.

In the context of a specific, longer computation, and depending on the parameter n and the target platform, we note that additional reductions might need to be applied to a limited number of intermediate values, for which overflow may occur. In this case, as an optimization, two successive reductions can be merged as follows. Let the input operand C be decomposed as $C = C_0 + C_1 \cdot 2^m + C_2 2^{2m}$ with $0 \le C_0, C_1 < 2^m$. Then we can reduce C via the following function `K-RED-2x`.

```
function K-RED-2x($C$)
    $C_0 \leftarrow C \bmod 2^m$
    $C_1 \leftarrow C/2^m \bmod 2^m$
    $C_2 \leftarrow C/2^{2m}$
    return $k^2 C_0 - kC_1 + C_2$
```

Speeding up the NTT. In the context of the NTT algorithm, we use a redundant representation of integers modulo q by allowing them to grow up to 32 bits and, when necessary, apply the reduction function `K-RED` to reduce the sizes of coefficients. We keep track of the factors of k that are implicitly multiplied to the result by an invocation of `K-RED`. For the sake of illustration, consider Algorithm 1. The main idea is to apply the function `K-RED` only after multiplications, i.e., one reduction per iteration in the inner loop, letting intermediate coefficient values grow such that the final coefficient values become congruent to $K \cdot a[\cdot] \bmod q$ for a fixed factor K. This factor can then be used at the end of the NTT-based polynomial multiplication to correct the result to the desired value. Next, we specify the details of the method for $n \in \{256, 512, 1024\}$ for the prime $q = 12289$. We limit the analysis to platforms with native 32 (or higher)-bit multipliers, but note that the presented algorithms can be easily modified to cover other settings.

The case $q = 12289$. The modified NTT algorithms using `K-RED` and `K-RED-2x` are shown in Algorithm 3 and Algorithm 4 for the modulus $q = 12289$, which in practice is used with $n = 512$ (for BLISS signatures [11]) or 1024 (for key exchange [2]). In Steps 7 of Algorithm 3 and Step 7 of Algorithm 4, we are using the precomputed values scaled by k^{-1}, i.e. we use precomputed tables $\Psi_{rev,k^{-1}}[\cdot] = k^{-1} \cdot \Psi_{rev}[\cdot]$ and $\Psi^{-1}_{rev,k^{-1}}[\cdot] = k^{-1} \cdot \Psi^{-1}_{rev}[\cdot]$. We denote these modified algorithms by $\texttt{NTT}^K := \mathrm{NTT}^K_{CT,\psi_{rev,k^{-1}}}$ and $\texttt{INTT}^K := \mathrm{INTT}^K_{GS,\Psi^{-1}_{rev,k^{-1}}}$, respectively.

Algorithm 3. Modified function $\texttt{NTT}^K$ using `K-RED` and `K-RED-2x` for reduction modulo $q = 12289$ (32 or 64-bit platform).

Input: A vector $a = (a[0], a[1], ..., a[n-1]) \in \mathbb{Z}_q^n$ in standard ordering, where $n \in \{256, 512, 1024\}$, and a precomputed table $\Psi_{rev,k^{-1}} \in \mathbb{Z}_q^n$ of scaled powers of ψ in bit-reversed order.
Output: $a \leftarrow \texttt{NTT}^K(a)$ in bit-reversed ordering.

```
 1: $t = n$
 2: for ($m = 1$; $m < n$; $m = 2m$) do
 3:     $t = t/2$
 4:     for ($i = 0$; $i < m$; $i{+}{+}$) do
 5:         $j_1 = 2 \cdot i \cdot t$
 6:         $j_2 = j_1 + t - 1$
 7:         $S = \Psi_{rev,k^{-1}}[m + i]$
 8:         for ($j = j_1$; $j \leq j_2$; $j{+}{+}$) do
 9:             $U = a[j]$
10:             $V = a[j + t] \cdot S$
11:             if $m = 128$ then
12:                 $U = \texttt{K-RED}(U)$
13:                 $V = \texttt{K-RED-2x}(V)$
14:             else
15:                 $V = \texttt{K-RED}(V)$
16:             $a[j] = U + V$
17:             $a[j + t] = U - V$
18: return $a$
```

Given two input vectors a and b, let $c = \texttt{INTT}(\texttt{NTT}(a) \circ \texttt{NTT}(b))$ be computed using Algorithms 1 and 2. It is easy to see that the resulting coefficients after applying Algorithms 3 and 4, i.e., after computing $\texttt{INTT}^K(\texttt{NTT}^K(a) \circ \texttt{NTT}^K(b))$, are congruent to $K \cdot c[\cdot]$ modulo q for a certain fixed integer $K = k^s$ and an integer s. Note that by scaling the precomputed twiddle factors by $k^{-1} \mod q$, we can limit the growth of the power of k introduced by the reduction steps.

Algorithm 4. Modified function $\texttt{INTT}^K$ using `K-RED` and `K-RED-2x` for reduction modulo $q = 12289$ (32 or 64-bit platform).

Input: A vector $a = (a[0], a[1], ..., a[n-1]) \in \mathbb{Z}_q^n$ in bit-reversed ordering, where $n \in \{256, 512, 1024\}$, a precomputed table $\Psi^{-1}_{rev,k^{-1}} \in \mathbb{Z}_q^n$ of scaled powers of ψ^{-1} in bit-reversed order, and constants $n_K^{-1} = n^{-1} \cdot k^{-11}$, $\Psi_K^{-1} = n^{-1} \cdot k^{-10} \cdot \Psi^{-1}_{rev,k^{-1}}[1] \in \mathbb{Z}_q$, where $k = 3$.
Output: $a \leftarrow \texttt{INTT}^K(a)$ in standard ordering.

```
 1: $t = 1$
 2: for ($m = n$; $m > 2$; $m = m/2$) do
 3:     $j_1 = 0$
 4:     $h = m/2$
 5:     for ($i = 0$; $i < h$; $i$++) do
 6:         $j_2 = j_1 + t - 1$
 7:         $S = \Psi^{-1}_{rev,k^{-1}}[h + i]$
 8:         for ($j = j_1$; $j \le j_2$; $j$++) do
 9:             $U = a[j]$
10:             $V = a[j + t]$
11:             $a[j] = U + V$
12:             $a[j + t] = (U - V) \cdot S$
13:             if $m = 32$ then
14:                 $a[j] = \texttt{K-RED}(a[j])$
15:                 $a[j + t] = \texttt{K-RED-2x}(a[j + t])$
16:             else
17:                 $a[j + t] = \texttt{K-RED}(a[j + t])$
18:         $j_1 = j_1 + 2t$
19:     $t = 2t$
20: for ($j = 0$; $j < t$; $j$++) do
21:     $U = a[j]$
22:     $V = a[j + t]$
23:     $a[j] = \texttt{K-RED}((U + V) \cdot n_K^{-1})$
24:     $a[j + t] = \texttt{K-RED}((U - V) \cdot \Psi_K^{-1})$
25: return $a$
```

For example in Line 7 of Algorithm 3 the value S carries a factor k^{-1} which then cancels with the factor k introduced by `K-RED` in Step 15. Only additional reductions such as those in Steps 12 and 13 increase the power of k in the final result.

At the end of the computation, the final results can be converted back to the standard representation by multiplying with the inverse of the factor K. Moreover, this conversion can be obtained for free if the computation is merged with the scaling by n^{-1} during the inverse transformation, that is, if scaling is performed by multiplying the resulting vector with the value $n^{-1} \cdot K^{-1}$. However, we can do even better: by merging the second entry of the table $\Psi_{rev,k^{-1}}$ with the fixed value $n^{-1} \cdot K^{-1}$, we eliminate an additional $n/2$ multiplications and modular reductions. This is shown in Steps 21–24 of Algorithm 4.

4 Case Study: R-LWE Key Exchange

This section explains how we apply our new modular reduction and the improved NTT algorithms, together with a simplified message encoding, to the key exchange implementation that was proposed by Alkim, Ducas, Pöppelmann and Schwabe in [2]; the protocol is depicted in Fig. 1. Accordingly, from hereon we fix $n = 1024$ and $q = 12289$ and the error distribution is defined to be the centered binomial distribution ψ_{12}, from which one samples by computing $\sum_{i=1}^{16}(b_i - b'_i)$, where the $b_i, b'_i \in \{0,1\}$ are uniform independent bits. The functions `HelpRec` and `Rec` are modified instantiations of Peikert's reconciliation functions [24, Sect. 3] that essentially turn approximate key agreement into *exact* key agreement – see [2]. The function `SHAKE-128` is the extended output function (XOF) based on Keccak [4], which is also used to derive the 256-bit shared secret key in both Alice and Bob's final steps. Following [2], the random value a is generated directly in the NTT domain.

Public parameters		
$n = 1024$, $q = 12289$, error distribution ψ_{12}		
Alice (server)		**Bob (client)**
$\mathit{seed} \xleftarrow{\$} \{0,1\}^{256}$		
$a \leftarrow \texttt{SHAKE-128}(\mathit{seed})$		
$s, e \xleftarrow{\$} \psi_{12}^n$		$s', e', e'' \xleftarrow{\$} \psi_{12}^n$
$b \leftarrow as + e$	$\xrightarrow{m_A=(b,\mathit{seed})}$	$a \leftarrow \texttt{SHAKE-128}(\mathit{seed})$
		$u \leftarrow as' + e'$
		$v \leftarrow bs' + e''$
$v' \leftarrow us$	$\xleftarrow{m_B=(u,r)}$	$r \xleftarrow{\$} \texttt{HelpRec}(v)$
$\nu \leftarrow \texttt{Rec}(v', r)$		$\nu \leftarrow \texttt{Rec}(v, r)$
$\mu \leftarrow \texttt{SHA3-256}(\nu)$		$\mu \leftarrow \texttt{SHA3-256}(\nu)$

Fig. 1. The key exchange instantiation from [2].

Viewing Fig. 1, we identify the following NTT-based computations:

Alice	Bob
$b \leftarrow a \circ \mathrm{NTT}(s) + \mathrm{NTT}(e)$	$u \leftarrow a \circ \mathrm{NTT}(s') + \mathrm{NTT}(e')$
$v' \leftarrow \mathrm{INTT}\,(u \circ \mathrm{NTT}(s))$	$v \leftarrow \mathrm{INTT}\,(b \circ \mathrm{NTT}(s') + \mathrm{NTT}(e''))$

The sequence of NTT and INTT operations above are used to determine the value of K that results from our target parameters; note that $q = 3 \cdot 2^{12} + 1$ and thus $k = 3$. For determining K, Alice's and Bob's NTT/INTT computations can be seen as *two* polynomial operations: (1) the first operation begins with the computation of b on Alice's side, who then transmits it in the NTT domain to Bob for computing v and giving the result back in the standard domain; and similarly (2) the second operation consists of the computation of u on Bob's side followed by the computation of v' on Alice's side.

We first point out that if we include two extra reductions at stage $m = 128$ and $m = 32$ of the NTT and INTT algorithms, respectively, then intermediate values never grow beyond 32 bits during a full NTT or INTT computation (see steps 11–13 of Algorithm 3 and steps 13–15 of Algorithm 4). Following Sect. 3, the factor k introduced by every invocation of `K-RED` is canceled out by the corresponding multiplication with an entry from the $\Psi_{\mathrm{rev},k^{-1}}$ and $\Psi^{-1}_{\mathrm{rev},k^{-1}}$ tables. Hence, only the extra reductions above introduce a factor k to the intermediate results of the NTT and INTT.

Secondly, we point out that after performing component-wise multiplications of polynomials in the NTT domain, the individual factors get compounded. The results after these multiplications require two additional reductions and a conditional subtraction per coefficient to fully reduce them modulo q (this is required to avoid overflows and, when applicable, to transmit messages and derive shared keys in fully reduced form). It is important to keep track of these factors and to (i) ensure that they are balanced (i.e., the same) before, e.g., adding two summands that are the result of different NTT operations, and (ii) ensure that they are corrected at the end of the computation. Careful analysis of the above sequence of NTT operations reveals that the final factor is $K = k^{10} = 3^{10}$ for the two full polynomial operations mentioned before.

Message Encoding and Decoding. Internally, polynomials are encoded as 1024-element little-endian arrays, where each element or coefficient is represented either by a 32-bit signed integer (for secret keys and error polynomials) or a 32-bit unsigned integer (for everything else). Each coefficient that is part of a message is fully reduced modulo q before transmission and therefore only uses a fraction of the integer size (i.e., 14 bits). We simply encode messages in little endian format as a concatenation of these 1024 14-bit coefficients (for b and u; see Fig. 1) immediately followed by the 256-bit *seed* in Alice's message and the 1024 2-bit array r in Bob's message. Accordingly, m_A and m_B consist of 1824 and 2048 bytes, respectively.

5 Implementation Results

In this section, we present implementation results showcasing the performance of the new NTT algorithms and, in particular, benchmark them in the context of the Ring-LWE key exchange by Alkim et al. [2].

5.1 Performance Benchmarks

To ease the comparison with the state-of-the-art NTT implementation, we followed [2] and implemented *two* versions of the proposed NTT algorithms [18]:

a portable and compact implementation written in the C language, and a high-speed implementation written in x64 assembly and exploiting AVX2 instructions. For the AVX2 implementation we decided to use vector integer instructions, which are easier to work with and, according to our theoretical analysis, are expected to provide similar performance to a version using vector floating-point instructions.

The benchmarking results of our implementations are shown at the top of Table 1. These results were obtained by running the implementations on a 3.4 GHz Intel Core i7-4770 Haswell processor with TurboBoost disabled. For compilation we used gcc v4.9.2 for the C implementation and clang v3.8.0 for the AVX2 implementation.

As one can see, for the C version, the new forward and inverse NTT implementations are 1.84 and 1.88 times faster than the corresponding implementations from Alkim et al. [2]. In contrast, for the AVX2 version, the new algorithms appear to be slightly slower. However, this direct comparison does not account for the additional benefits of our technique that are not observable at the NTT level. This includes the efficient use of signed arithmetic and the elimination of costly conversion routines required by the Montgomery arithmetic (as used in [2]) that are performed outside of the NTT. As we show below, our algorithms perform significantly better in practice when all this additional overhead is considered in the cost.

Table 1. Benchmarking results (in terms of 10^3 cycles) of our C and AVX2 implementations of the NTT and the key-exchange instantiation proposed by Alkim et al. [2] on a 3.4 GHz Intel Core i7-4770 Haswell processor with TurboBoost disabled. Results are compared with Alkim et al.'s implementation results. At the bottom of the table, we show the total cost of a key-exchange, including Alice's and Bob's computations.

	C implementation		AVX2 implementation	
	ADPS [2]	This work	ADPS [2]	This work
NTT	55.4	30.1	8.4	9.1
INTT	59.9	31.8	9.5	9.7
Generating a	43.6	39.5	36.9	37.8
Error sampling	32.7	31.4	5.9	4.8
`HelpRec`	14.6	12.9	3.4	2.4
`Rec`	10.1	7.2	2.8	1.2
Key gen (server)	259.0	**170.9**	89.1	**70.4**
Key gen + shared key (client)	385.1	**287.6**	111.2	**95.2**
Shared key (server)	86.3	**48.8**	19.4	**15.7**
Total (key exchange)	730.4	**507.3**	219.7	**181.3**

To illustrate the *overall* performance benefits of the new reduction and NTT algorithms, we implemented the full key-exchange instantiation proposed by Alkim et al. [2]. To ease the comparison, we reuse the same implementations of ChaCha20 and `SHAKE-128` used in Alkim et al.'s software for the seed expansion during the generation of a and for the polynomial error sampling, respectively.

Our results for the key exchange are summarized in Table 1. The C and AVX2 implementations are roughly 1.44x and 1.21x faster, respectively, than the corresponding C and AVX2 implementations by Alkim et al. These improvements are mostly due to the new NTT algorithms which exhibit a faster reduction and avoid the costly conversions that are required when working with Montgomery arithmetic. The new reduction also motivates the use of signed arithmetic, which makes computations more efficient because corrections from negative to positive values are not required in several of the key exchange routines. In particular, the effect of using signed arithmetic can be observed in the performance improvement for the generation of a, `HelpRec` and `Rec`. We remark that these performance improvements are obtained with significantly simpler integer arithmetic.

A different Ring-LWE based key-exchange implementation has been recently reported by Aguilar-Melchor et al. [1]. Direct comparisons with this work are especially difficult because they use different parameters and the most recent version of their implementation appears not to be protected against timing and cache attacks. As a point of reference, we mention that [1, Table 2] reports that their NTT implementation using $n = 512$ and a 30-bit modulus runs in 13 K cycles on a 2.9 GHz Intel Haswell machine (scaled from 4.5 μs). This is more than 1.4x slower than our NTT using $n = 1024$ and a 14-bit modulus.

6 Conclusion

We describe a new modular reduction technique and improved FFT algorithms to compute the NTT. The improved NTT algorithms were applied to a recent key exchange proposal and showed significant improvements in performance using both a plain C implementation and a vectorized implementation that does not require floating-point arithmetic.

Although both the modular reduction and the improved NTT were motivated by (and are somewhat tailored towards) applications in R-LWE cryptography that use power-of-2 cyclotomic fields, our improvements should be of independent interest and might be applicable to other scenarios. Our method offers flexibility for implementations with different design goals without sacrificing performance.

Likewise, we expect that the new algorithms offer similar performance improvements on platforms such as microcontrollers and ARM processors. We leave this as future work, as well as the evaluation of the proposed NTT algorithms in the implementation and optimization of R-LWE signature schemes such as BLISS.

References

1. Aguilar-Melchor, C., Barrier, J., Guelton, S., Guinet, A., Killijian, M.-O., Lepoint, T.: NFLlib: NTT-based fast lattice library. In: Sako, K. (ed.) CT-RSA 2016. LNCS, vol. 9610, pp. 341–356. Springer, Heidelberg (2016). doi:10.1007/978-3-319-29485-8_20
2. Alkim, E., Ducas, L., Pöppelmann, T., Schwabe, P.: Post-quantum key exchange - a new hope. In: Holz, T., Savage, S. (eds.) 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, pp. 327–343. USENIX Association, 10–12 August 2016
3. Bai, S., Langlois, A., Lepoint, T., Stehlé, D., Steinfeld, R.: Improved security proofs in lattice-based cryptography: using the Rényi divergence rather than the statistical distance. In: Iwata, T., Cheon, J.H. (eds.) ASIACRYPT 2015. LNCS, vol. 9452, pp. 3–24. Springer, Heidelberg (2015). doi:10.1007/978-3-662-48797-6_1
4. Bertoni, G., Daemen, J., Peeters, M., Van Assche, G.: Keccak. In: Johansson, T., Nguyen, P.Q. (eds.) EUROCRYPT 2013. LNCS, vol. 7881, pp. 313–314. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38348-9_19
5. Bos, J.W., Costello, C., Naehrig, M., Stebila, D.: Post-quantum key exchange for the TLS protocol from the ring learning with errors problem. In: 2015 IEEE Symposium on Security and Privacy, SP 2015, pp. 553–570. IEEE Computer Society (2015)
6. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. TOCT **6**(3), 13:1–13:36 (2014)
7. Chu, E., George, A.: Inside the FFT Black Box Serial and Parallel Fast Fourier Transform Algorithms. CRC Press, Boca Raton (2000)
8. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Math. Comput. **19**(90), 297–301 (1965)
9. Crandall, R., Pomerance, C.: Prime Numbers: A Computational Perspective. Springer, Heidelberg (2005)
10. Ding, J., Xie, X., Lin, X.: A simple provably secure key exchange scheme based on the learning with errors problem. Cryptology ePrint Archive, Report 2012/688 (2012). http://eprint.iacr.org/2012/688
11. Ducas, L., Durmus, A., Lepoint, T., Lyubashevsky, V.: Lattice signatures and bimodal Gaussians. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013. LNCS, vol. 8042, pp. 40–56. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40041-4_3
12. Gauss, C.F.: Nachlass, theoria interpolationis methodo nova tractata. In: Carl Friedrich Gauss Werke, Band 3, pp. 265–330 (1866)
13. Gentleman, W.M., Sande, G.: Fast, fourier transforms: for fun and profit. In: Fall Joint Computer Conference, AFIPS 1966, pp. 563–578, ACM, New York (1966)
14. Harvey, D.: Faster arithmetic for number-theoretic transforms. J. Symb. Comput. **60**, 113–119 (2014)
15. Hoffstein, J., Pipher, J., Silverman, J.H.: NTRU: a ring-based public key cryptosystem. In: Buhler, J.P. (ed.) ANTS 1998. LNCS, vol. 1423, pp. 267–288. Springer, Heidelberg (1998). doi:10.1007/BFb0054868
16. Knuth, D.E.: Seminumerical algorithms. In: Lai, V.S., Mahapatra, R.K. (eds.) The Art of Computer Programming, 3rd edn. Addison-Wesley, Reading (1997)
17. Liu, Z., Seo, H., Roy, S.S., Großschädl, J., Kim, H., Verbauwhede, I.: Efficient ring-LWE encryption on 8-Bit AVR processors. In: Güneysu, T., Handschuh, H. (eds.) CHES 2015. LNCS, vol. 9293, pp. 663–682. Springer, Heidelberg (2015). doi:10.1007/978-3-662-48324-4_33

18. Longa, P., Naehrig, M.: LatticeCrypto (2016). https://www.microsoft.com/en-us/research/project/lattice-cryptography-library/
19. Lyubashevsky, V.: Lattice signatures without trapdoors. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 738–755. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29011-4_43
20. Lyubashevsky, V., Peikert, C., Regev, O.: On ideal lattices and learning with errors over rings. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 1–23. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13190-5_1
21. Lyubashevsky, V., Micciancio, D., Peikert, C., Rosen, A.: SWIFFT: a modest proposal for FFT hashing. In: Nyberg, K. (ed.) FSE 2008. LNCS, vol. 5086, pp. 54–72. Springer, Heidelberg (2008). doi:10.1007/978-3-540-71039-4_4
22. Micciancio, D., Peikert, C.: Hardness of SIS and LWE with small parameters. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013. LNCS, vol. 8042, pp. 21–39. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40041-4_2
23. Nussbaumer, H.J.: Fast polynomial transform algorithms for digital convolution. IEEE Trans. Acoust. Speech Sig. Process. **28**(2), 205–215 (1980)
24. Peikert, C.: Lattice cryptography for the internet. In: Mosca, M. (ed.) PQCrypto 2014. LNCS, vol. 8772, pp. 197–219. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11659-4_12
25. Pöppelmann, T., Güneysu, T.: Towards practical lattice-based public-key encryption on reconfigurable hardware. In: Lange, T., Lauter, K., Lisoněk, P. (eds.) SAC 2013. LNCS, vol. 8282, pp. 68–85. Springer, Heidelberg (2014). doi:10.1007/978-3-662-43414-7_4
26. Pöppelmann, T., Oder, T., Güneysu, T.: High-performance ideal lattice-based cryptography on 8-bit ATXmega microcontrollers. In: Lauter, K., Rodríguez-Henríquez, F. (eds.) LATINCRYPT 2015. LNCS, vol. 9230, pp. 346–365. Springer, Heidelberg (2015). doi:10.1007/978-3-319-22174-8_19
27. Proth, F.: Théorèmes sur les nombres premiers. Comptes Rendus des Séances de l'Académie des Sciences, Paris **87**, 926 (1878)
28. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. In: Proceedings of the 37th Annual ACM Symposium on Theory of Computing, pp. 84–93 (2005)
29. Roy, S.S., Vercauteren, F., Mentens, N., Chen, D.D., Verbauwhede, I.: Compact ring-LWE cryptoprocessor. In: Batina, L., Robshaw, M. (eds.) CHES 2014. LNCS, vol. 8731, pp. 371–391. Springer, Heidelberg (2014). doi:10.1007/978-3-662-44709-3_21
30. Shoup, V.: Number Theory Library (NTL), 1996–2016. http://www.shoup.net/ntl
31. Stehlé, D., Steinfeld, R.: Making NTRU as secure as worst-case problems over ideal lattices. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 27–47. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20465-4_4
32. Zhang, J., Zhang, Z., Ding, J., Snook, M., Dagdelen, Ö.: Authenticated key exchange from ideal lattices. In: Oswald, E., Fischlin, M. (eds.) EUROCRYPT 2015. LNCS, vol. 9057, pp. 719–751. Springer, Heidelberg (2015). doi:10.1007/978-3-662-46803-6_24