Accelerating integer-based fully homomorphic encryption using Comba multiplication

https://doi.org/10.1109/SiPS.2014.6986063

Published in:
Proceedings of 2014 IEEE Workshop on Signal Processing Systems (SiPS)

Document Version:
Early version, also known as pre-print

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Accelerating Integer-based Fully Homomorphic Encryption using Comba Multiplication

Ciara Moore, Máire O’Neill, Neil Hanley, Elizabeth O’Sullivan
Centre for Secure Information Technologies
Queen’s University Belfast
{cmoore50, maire.oneill, n.hanley, e.osullivan}@qub.ac.uk

Abstract—Fully Homomorphic Encryption (FHE) is a recently developed cryptographic technique which allows computations on encrypted data. There are many interesting applications for this encryption method, especially within cloud computing. However, the computational complexity is such that it is not yet practical for real-time applications. This work proposes optimised hardware architectures of the encryption step of an integer-based FHE scheme with the aim of improving its practicality. A low-area design and a high-speed parallel design are proposed and implemented on a Xilinx Virtex-7 FPGA, targeting the available DSP slices, which offer high-speed multiplication and accumulation. Both use the Comba multiplication scheduling method to manage the large multiplications required with uneven sized multiplicands and to minimise the number of read and write operations to RAM. Results show that speed up factors of 3.6 and 10.4 can be achieved for the encryption step with medium-sized security parameters for the low-area and parallel designs respectively, compared to the benchmark software implementation on an Intel Core2 Duo E8400 platform running at 3 GHz.

I. INTRODUCTION

Fully Homomorphic Encryption (FHE) is a recently developed type of encryption scheme, introduced by Gentry [1], which enables computations on data while it remains in an encrypted form. A somewhat homomorphic encryption (SHE) scheme is firstly created, which allows a limited number of additions and multiplications of ciphertexts. This is extended using the techniques, such as squashing and bootstrapping, proposed by Gentry in [1] to create a FHE scheme, which supports unlimited multiplications and additions of ciphertexts. Random noise is generated with each operation, in particular with multiplications, and there are various methods proposed to manage this noise as it grows with each computation, such as modulus switching, which has been proposed by Brakerski et al. [2]. An application of FHE, for example, is secure computation on the cloud whereby users could take advantage of the cloud computing platform, without having to disclose data to a third party service provider. Another application of SHE and FHE is within multi-party computation, and schemes for this purpose have been proposed [3], [4].

FHE schemes have greatly advanced over recent years, since its introduction in 2009 [1]. There has been a vast array of work in the theoretical domain [2], [5]–[13], however current schemes are not yet efficient enough for real-time applications. For example, an evaluation of AES using FHE is reported to take around 36 hours on a machine with 256 GB RAM [10]; in another software implementation of the FHE scheme proposed by Gentry and Halevi, bitwise encryption at the highest security level is stated to take 3 minutes and the public key sizes required in this scheme range from 17 MB to 2.25 GB [5].

Thus, to address this shortcoming, optimised architectures that target alternative platforms, such as Graphics Processing Units (GPUs) and FPGAs, have recently been proposed [14]–[19]. Wang et al. [14] implemented Gentry and Halevi’s FHE scheme [5] on the NVIDIA C2050 GPU and achieved speed up factors of around 7 compared to the original implementation at a small security level. Furthermore, a design for a large-number multiplier for FHE targeting a Stratix-V FPGA technology was presented last year by Wang et al. [15]; this multiplier uses the FFT algorithm and is reported to be twice as fast as the same multiplication implemented on the NVIDIA C2050 GPU.

Previously, Cao et al. proposed the first hardware implementation [16] of the encryption step of an FHE scheme over the integers [11], specifically investigating the use of a Virtex-7 FPGA platform. The experimental results reported a speed improvement factor of 11.25 for an implementation on a Virtex-7 XC7VX980T FPGA of the encryption step using the large security parameters compared to the original software implementation [11]. An extended version of this work is available on the IACR ePrint Archive [20] and includes optimised hardware architectures for the encryption step of the two integer-based FHE schemes [11], [12].

Architectures targeting Application Specific Integrated Circuit (ASIC) technology have also been proposed to improve the performance of FHE schemes [17], [18]. Doröz et al presented a custom hardware architecture for a million-bit multiplier for the implementation of Gentry and Halevi’s FHE scheme [5], and estimates show similar performance to the original software implementation. All of the previous hardware and GPU implementations to date employ the Fast Fourier Transform (FFT) to perform the large scale multiplications required for these FHE schemes.

The objective of this work is to design an optimised architecture for the implementation of the encryption step of the FHE scheme over the Integers specifically tailored to a FPGA device. Unlike previous work which used the FFT for the large multiplication operations, an alternative method for large integer multiplication is used in our work, which
utilises the high speed DSP multiplication blocks (DSP48E1s) available on Xilinx Virtex-7 FGPA to implement the Comba multiplication scheduling method [21]. This builds on previous work on the use of DSP slices and Comba multiplication for the hardware acceleration of FHE [19], where a hardware architecture for a Comba multiplier was proposed and estimated timings were given for the large integer multiplier required in the encryption step of an integer-based encryption scheme [12]. Rather than estimating timings, this work presents implementation and synthesis results for a Xilinx Virtex-7 FPGA of two optimised hardware architectures for the encryption step.

A Xilinx Virtex-7 FPGA is targeted in particular in this work because of the high suitability of FPGAs for DSP applications. Virtex-7 FGPA contain many DSP slices, each offering dedicated (25 x 18)-bit multiplication and 48-bit accumulation, which can run at frequencies of up to 741 MHz [22]. Moreover FGPA are reconfigurable which allows for fast prototyping and testing.

There are several types of FHE schemes, which have developed from the original lattice-based schemes proposed by Gentry [1], [5], [6]. More recent schemes are based on the learning with errors and ring learning with errors problems, by Gentry [1], [5], [6]. More recent schemes are based on the developments in the theoretical domain, which have improved the performance by minimising the size of the public keys required [12] and with the use of batching techniques [13]. Moreover, the designs proposed include building blocks, such as large integer multiplication, which are integral to most FHE schemes, such as Gentry and Halevi’s FHE scheme [5].

The remainder of this paper is organised as follows: Section II gives an overview of FHE over the integers with particular focus on the encryption step to be implemented; the Comba multiplication technique is outlined in Section III; Section IV details the Barrett modular reduction method used in the implementations; in Sections V and VI the two novel architectures for the FHE encryption step are presented; synthesis results are detailed in Section VI, along with a comparison to previous implementations and finally, Section VII concludes the paper.

II. FULLY HOMOMORPHIC ENCRYPTION OVER THE INTEGERS

FHE over the Integers was introduced in 2010 by van Dijk et al. [7]. This type of FHE scheme is relatively simpler than other FHE schemes and is based on the Approximate GCD problem: given several $x_i$, where $x_i = p \cdot q_i + r_i$, find the secret key $p$. The original FHE scheme over the integers [7] was subsequently extended by Coron et al [11], [12], where in particular the public key sizes were reduced, through the use of pseudo random number generation.

This work focuses on the encryption step in the scheme proposed in [12]. The encryption step is one of several steps within the scheme and contains two important building blocks, large integer multiplication and modular reduction. Although we initially focus solely on the encryption step in this work, the construction and optimisation of these building blocks required in the encryption step will also be used in future work to implement the other steps within this FHE scheme, and moreover can also be used to implement other FHE schemes.

The encryption step is defined as

$$c \leftarrow m + 2r + 2 \sum_{i=1}^{\tau} x_i \cdot b_i \mod x_0 \quad (1)$$

where $m \in \{0,1\}$ is the message bit; $r$ is a random noise parameter; $x_i$ are integers generated from the public key as described in the key generation step in [12]; $b_i$ are randomly selected integers, which are much smaller than the $x_i$. The parameter sizes are given in Table I. The selection of suitable parameters is out of the scope of this current work; for more information on the security levels, parameter selection and for detail on the rest of the FHE scheme, see the original work by van Dijk et al [7] and Coron et al [11], [12].

<table>
<thead>
<tr>
<th>Param. Sizes</th>
<th>Bit-length, $x_i$</th>
<th>Bit-length, $b_i$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy</td>
<td>150k</td>
<td>936</td>
<td>158</td>
</tr>
<tr>
<td>Small</td>
<td>830k</td>
<td>1476</td>
<td>572</td>
</tr>
<tr>
<td>Medium</td>
<td>4.2m</td>
<td>2016</td>
<td>2110</td>
</tr>
<tr>
<td>Large</td>
<td>19.0m</td>
<td>2556</td>
<td>7695</td>
</tr>
</tbody>
</table>

III. COMBA MULTIPLICATION

As can be seen from Equation (1), a large multiply-accumulate is required, which is the main bottleneck in the encryption step. As mentioned in Section II, multiplication is also needed in other steps and in other FHE schemes in the literature, hence the work is transferable. Currently, the approach taken by the research community is to use the Fast Fourier Transform for fast multiplication, which is suitable for very large multiplication sizes.

An alternative fast multiplication method is Karatsuba multiplication [23], which is asymptotically faster than traditional schoolbook multiplication. It requires intermediate values to be stored for each multiplication, and hence is not very suitable for an FPGA implementation of these FHE schemes, as they require very large multiplications and the storage of intermediate values would be problematic.

Therefore, we take an alternative approach and make use of the fast embedded multiplication blocks available within the DSP slices on Xilinx Virtex-7 FGPA and combine this with the Comba scheduling method [21]. As can be seen from Table I, the $x_i$ are much greater than the $b_i$. To minimise the number of read and write operations, the multiplication block width $w$ is set to the next power of two greater than or equal to the the $b_i$ bit length. Thus, the multiplication block width
is 1024, 2048, 4096 and 4096 for the four parameter groups respectively.

Comba multiplication is a scheduling method to effectively control the multiplication and accumulation of partial products [21]. It reduces the number of expensive write accesses to memory compared to traditional school book multiplication. For example, if the Comba scheduling method is used to multiply two large integers \(x\) and \(y\), these integers are divided into several smaller words; thus \(x\) and \(y\) have \(n_0\) and \(n_1\) words respectively, and these are multiplied and accumulated to calculate the large integer multiplication of \(x\) and \(y\). Instead of writing the \(n_0 \times n_1\) partial products to memory, as necessary for schoolbook multiplication, only \(n_0 + n_1 - 1\) partial products are required when the Comba scheduling method is used. Following Algorithm 1, after each partial product multiplication-accumulation step the least significant word is written to memory and the remainder is shifted and then accumulated with the next partial product.

Algorithm 1: Comba partial product accumulation

```plaintext
Input: \(n_0\)-word \(x\), \(n_1\)-word \(y\), where \(n_0 \leq n_1\)
Output: \((n_0 + n_1 - 1)\) partial products, \(pp_i\)
1: for \(i\) in 0 to \((n_0 + n_1 - 2)\) do
2: if \(n_0 < i\) then
3: \(pp_i = \sum_{k=0}^{i-1} (x_k \times b_{i-k})\)
4: else
5: \(pp_i = \sum_{k=0}^{n_1-1} (x_k \times b_{i-k})\)
6: end if
7: end for
return \(pp_i\)
```

Güneysu demonstrates a parallelised Comba multiplication method, which takes advantage of the available DSP blocks on an FPGA [24]. The accumulation within the Comba multiplication is also carried out using the dedicated accumulator available in each DSP slice. We take a similar approach for the multiplications required in Equation (1).

IV. BARRET Modular Reduction

The most commonly used modular reduction methods are the Montgomery and Barrett methods. Modular reduction involves division and hence is a costly operation in hardware. We use an improved Barrett reduction method in our design rather than Montgomery, which requires expensive pre-processing and post-processing to and from the Montgomery domain. This lends itself more to repeated reduction operations, such as in an exponentiation operation, which Equation (1) does not require.

Barrett reduction is used to carry out the modular reduction required in the encryption step in Equation (1). The reduction algorithm uses two multiplications and a subtraction. The same multiplication block will be used for both the multiplication and the modular reduction in the proposed low-area design, as these operations are sequential. This minimises the hardware area usage.

The Barrett Reduction method is optimised, as proposed by Dhem [25], so that only one subtraction is needed after the multiplication, when \(\alpha \geq m\) and \(\beta \leq -2\), as described in Algorithm 2. This algorithm was also previously used in the FPGA implementation of the encryption step of integer-based FHE [16]; however as mentioned, this previous work used the FFT algorithm rather than the Comba multiplication algorithm as used here.

Algorithm 2: Barrett reduction [25], [26]

```plaintext
Input: \(x, m\)-bit \(p, \alpha, \beta\) and a precomputed constant \(p_1 = \left[\frac{2^{m+\alpha}}{p}\right]\)
Output: \(y = x \mod p\)
1: \(\sigma = \left[\frac{x/2^{n-p}}{2^{n-p}p_1}\right];\)
2: \(\hat{p} = \sigma \times p;\)
3: \(y_1 = x - \hat{p}\) and \(y_2 = y_1 - \hat{p};\)
4: if \(y_2 < 0\) then
5: \(y = y_1\)
6: else
7: \(y = y_2\)
8: end if
return \(y\)
```

V. HARDWARE ARCHITECTURE FOR ENCRYPTION STEP

The main building block in the encryption step (1) to be implemented is the multiplication-accumulation unit. Two architectures are proposed; both architectures contain a finite state machine, controlling the data unit and the read and write operations. The designs both use off-chip memory to store input operands, intermediate values and final results. The main difference between the two architectures is the number of multiplication blocks: the first design contains just one multiplication block, minimising resource usage, and the second design contains several multiply-accumulate blocks, maximising speed.

The DSP slices available on Xilinx Virtex-7 FPGAs are used within the multiplication blocks in both designs. Figure 1 is a structural overview of a DSP slice, as given in the Xilinx 7 Series DSP48E1 Slice User Guide [27]. Both a 25 × 18-bit signed multiplier and a 48-bit accumulation unit are offered in each slice. Thus, a 16-bit unsigned multiplier is used in each slice in the multiplication blocks in both designs in this section. Each block consists of \(\frac{w}{16}\) DSP slices, where \(w\) is the width of the overall multiplication in each block and is given in Table II.

The target of this work is an optimised architecture and implementation of the encryption step introduced in Section II, in order to improve its performance. Thus, it is assumed that there is enough off-chip memory available to store input operands, intermediate values and final results. This assumption is reasonable, as the designs proposed in this work use 64-bit read and write operations and the FPGA can access shared memory, for example with a PC, using a high speed PCI bus.
A. Low-area Architecture Design

Figure 2 depicts the architecture of the first proposed design, which aims to minimise the area resource required. The main element in this architecture is one multiplication block of width \( w \), approximately the size of the smaller multiplicand \( b_i \), which is controlled by a finite state machine. The same multiplication block is also repeatedly used within the modular reduction step. The design therefore minimises area usage, but obviously this somewhat limits the performance. The multiplexers control the adders for two uses: addition required for the multiplication-accumulation operation (when \( acc\_sel = 1 \)) and secondly subtraction required in the modular reduction operation (when \( acc\_sel = 0 \)).

B. Parallel Architecture Design

The goal of the second design is speed. The maximum possible number of multiply-accumulate blocks that can fit on a high-end Virtex-7 FPGA are used in this architecture in order to maximise the performance of the encryption step. The Virtex-7 FPGA XC7VX980T is targeted because it has the largest number of DSP slices and a large amount of logic cells. The parallel architecture is outlined in Figure 4 and uses multiple multiply-accumulate blocks. The architecture of an individual multiply-accumulate block is illustrated in Figure 3. The number of blocks depends on the size of the multiplier \( w \), listed in Table II, which is the size of next power of two greater than or equal to the number of bits in the smaller operand in the multiplication block \( b_i \). As the size of the \( b_i \) operand increases with increasing parameter security levels, from the smallest toy security setting to the large security parameter setting, the number of multiplication blocks that can fit on the FPGA decreases.

VI. Results

The two proposed architectures were implemented using the Xilinx ISE Design Suite 14.1 synthesis tool. The target device is the Virtex-7 XC7VX980T-2FFG1926, the Virtex-7 FPGA with the largest number of DSP slices. The synthesis results and hardware area usage are given in Table II and Table III for the parallel and low-area designs respectively. These designs fit comfortably on the target FPGA device. Moreover, the low-area design achieves high clock frequencies of around 300 MHz for the four parameter security sizes, which are much greater than the clock frequencies achieved in the implementation of the high-speed architecture. It should be noted that post place and route of the designs will give slightly worse results; however the speed of the read and write operations can be greatly improved upon, with the use of external memory interfaces and for example DDR3 SDRAM memory, which can run at four times the clock frequency [28].

As expected, the hardware resource utilisation is much greater for the implementation of the encryption step using
TABLE II: Synthesis results for the high-speed parallel architecture

<table>
<thead>
<tr>
<th>Param. Sizes</th>
<th>$w$, Multiplication block bit-width</th>
<th>Parallel Mul-Acc Blocks</th>
<th>Freq. (MHz)</th>
<th>DSP48E1s</th>
<th>Slice Reg</th>
<th>Slice LUTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy</td>
<td>1024</td>
<td>20</td>
<td>197.465</td>
<td>1344</td>
<td>469577</td>
<td>262054</td>
</tr>
<tr>
<td>Small</td>
<td>2048</td>
<td>10</td>
<td>192.806</td>
<td>1408</td>
<td>491712</td>
<td>311914</td>
</tr>
<tr>
<td>Medium</td>
<td>2048</td>
<td>10</td>
<td>197.758</td>
<td>1408</td>
<td>497773</td>
<td>313104</td>
</tr>
<tr>
<td>Large</td>
<td>4096</td>
<td>5</td>
<td>197.758</td>
<td>1536</td>
<td>542949</td>
<td>365315</td>
</tr>
</tbody>
</table>

The timings in Table IV are calculated using the clock cycle count and the synthesized design frequencies are stated in Table II and Table III. The read and write operations to external RAM are included within the clock cycle count; two clock cycles are used to read or write each 64-bit word to and from the memory. For example, the clock cycle latency for the multiplication operation is calculated as follows: the multiplication unit requires $2 \times \lceil \frac{w}{16} \rceil + 2$ clock cycles, where 16-bits is the size of the multiplication used within each DSP block and $w$ is the multiplication block width, defined in Table II. In the first architecture, $\tau \times \lceil |x_i|/w \rceil$ multiplications are required. Thus for solely the multiplication operation required in the low-area design, for example the toy sized design requires $158 \times \lceil \frac{150000}{1024} \rceil \times (2 \times \lceil \frac{1024}{16} \rceil + 2) = 3019380$ clock cycles, which takes around 0.0098s.

As can be seen in Table IV, both proposed architectures perform better than the benchmark software implementation [20]. The first and second designs are approximately 3 times and 8-10 times faster respectively when compared to the original results. This matches the speed up factors achieved by other hardware implementations of alternative FHE schemes.

Cao et al. [20] have also implemented optimised architectures for the same encryption step (1) targeting the same Virtex-7 FPGA device and using FFT for the large integer multiplication. For comparison purposes, the timing results of the design which fits on the target Virtex-7 FPGA [20] are given in the last row of Table IV. As can be seen from this table, our proposed high-speed architecture performs consistently better than the previous implementations. For example for the medium parameter size, the high-speed architecture is 10.5 times faster than the original software implementation and 3.8 times faster than the implementation using the FFT multiplier [20]. Moreover, our design reduces the number of write operations required by using the Comba scheduling method instead of the schoolbook multiplication scheduling used in [20], which requires intermediate read and write operations. The Comba scheduling method employed in the proposed designs in this work ensures that the partial products are accumulated within the multiply-accumulate blocks, and the only write operations are for the digits of the final result of the multiply-accumulate operation, stated in Equation (1). To the best of the authors’ knowledge, no other comparable hardware architectures exist other than those provided in Table IV.

Our designs also use a reasonably small (64-bit) memory interface. These results from Table IV can also be further improved upon, for example, by using a pseudo-random number generator to generate public key values on the fly as

Fig. 4: Parallel architecture design of encryption step
proposed in [12], rather than reading from off-chip memory. This, and both algorithmic and implementation optimisations, such as batching, would further increase the practicality of these schemes.

VII. CONCLUSION

In conclusion, novel optimised low-area and parallel architectures of the encryption step of an integer-based FHE scheme are proposed and optimally implemented on a Virtex-7 FPGA, achieving speed up factors of around 3 and 10 when compared to the benchmark software implementation respectively. The results achieved illustrate that the use of an optimised architecture to a specific FPGA device which contains specialised DSP slices improves the practicality of the implementation of these FHE schemes and brings them closer to deployment. However further research into optimisations is still required, both at the algorithmic and the architectural level, before FHE schemes can be used in real-time applications. Future work will investigate the use of the Comba scheduling method with the FFT multiplier in order to provide further performance improvements to FHE.

REFERENCES


1Note, the integer-based FHE scheme by Coron et al. [12] is implemented on a single core of an Intel core2 Duo E8400 platform at 3 GHz.