1

HIGH PERFORMANCE ELLIPTIC CURVE CRYPTOGRAPHIC CO-PROCESSOR

Jonathan Lutz
General Dynamics - C4 Systems
Scottsdale, Arizona
E-mail: Jonathan.Lutz@gdc4s.com

M. Anwarul Hasan
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, ON, Canada
E-mail: ahasan@ece.uwaterloo.ca

For an equivalent level of security, elliptic curve cryptography uses shorter key sizes and is considered to be an excellent candidate for constrained environments like wireless/mobile communications. In FIPS 186-2, NIST recommends several finite fields to be used in the elliptic curve digital signature algorithm (ECDSA). Of the ten recommended finite fields, five are binary extension fields with degrees ranging from 163 to 571. The fundamental building block of the ECDSA, like any ECC based protocol, is elliptic curve scalar multiplication. This operation is also the most computationally intensive. In many situations it may be desirable to accelerate the elliptic curve scalar multiplication with specialized hardware.

In this chapter a high performance elliptic curve processor is described which is optimized for the NIST binary fields. The architecture is built from the bottom up starting with the field arithmetic units. The architecture uses a field multiplier capable of performing a field multiplication over the extension field with degree 163 in 0.060 microseconds. Architectures for squaring and inversion are also presented. The co-processor uses Lopez and Dahab's projective coordinate system and is optimized specifically for Koblitz curves. A prototype of the processor has been implemented for the binary extension field with degree 163 on a Xilinx XCV2000E FPGA. The prototype runs at 66 MHz and performs an elliptic curve scalar multiplication in 0.233 msec on a generic curve and 0.075 msec on a Koblitz curve.

1. INTRODUCTION

The use of elliptic curves in cryptographic applications was first proposed independently in [15] and [23]. Since then several algorithms have been developed whose

strength relies on the difficulty of the discrete logarithm problem over a group of elliptic curve points. Prominent examples include the Elliptic Curve Digital Signature Algorithm (ECDSA) [24], EC El-Gammal and EC Diffie Hellman [12]. In each case the underlying cryptographic primitive is elliptic curve *scalar* multiplication. This operation is by far the most computationally intensive step in each algorithm. In applications where many clients authenticate to a single server (such as a server supporting SSL [7, 26] or WTLS [1]), the computation of the scalar multiplication becomes the bottle neck which limits throughput. In a scenario such as this it may be desirable to accelerate the elliptic curve scalar multiplication with specialized hardware. In doing so, the scalar multiplications are completed more quickly and the computational burden on the server's main processor is reduced.

The selection of the ECC parameters is not a trivial process and, if chosen incorrectly, may lead to an insecure system [12, 24, 22]. In response to this issue NIST recommends ten finite fields, five of which are binary fields, for use in the ECDSA [24]. The binary fields include GF(2^{163}), GF(2^{233}), GF(2^{283}), GF(2^{409}) and GF(2^{571}) defined by the reduction polynomials in Table 1. For each field a specific curve, along with

Table 1. NIST Recommended Finite Fields

Field	Reduction Polynomial
GF(2^{163})	$F(x) = x^{163} + x^7 + x^6 + x^3 + 1$
GF(2^{233})	$F(x) = x^{233} + x^{74} + 1$
GF(2^{283})	$F(x) = x^{283} + x^{12} + x^7 + x^5 + 1$
GF(2^{409})	$F(x) = x^{409} + x^{87} + 1$
GF(2^{571})	$F(x) = x^{571} + x^{10} + x^5 + x^2 + 1$

a method for generating a pseudo-random curve, are supplied. These curves have been intentionally selected for both cryptographic strength and efficient implementation.

Such a recommendation has significant implications on design choices made while implementing elliptic curve cryptographic functions. In standardizing specific fields for use in elliptic curve cryptography (ECC), NIST allows ECC implementations to be heavily optimized for curves over a single finite field. As a result, performance of the algorithm can be maximized and resource utilization, whether it be in code size for software or logic gates for hardware, can be minimized.

Described in this chapter are hardware architectures for multiplication, squaring and inversion over binary finite fields. Each of these architectures is optimized for a

specific finite field with the intent that it might be implemented for any of the five NIST recommended binary curves. These finite field arithmetic units are then integrated together along with control logic to create an elliptic curve cryptographic co-processor capable of computing the scalar multiple of an elliptic curve point. While the co-processor supports all curves over a single binary field, it is optimized for the special Koblitz curves [16].

To demonstrate the feasibility and efficiency of both the finite field arithmetic units and the elliptic curve cryptographic co-processor, the latter has been implemented in hardware using a field programmable gate array (FPGA). The design was synthesized, timed and then demonstrated on a physical board holding an FPGA.

This chapter is organized as follows. Section 2 gives an overview of the basic mathematical concepts used in elliptic curve cryptography. This section also provides an introduction to the hardware/software system used to implement the elliptic curve scalar multiplier. Section 3 presents efficient hardware architectures for finite field multiplication and squaring. A method for high speed inversion is also discussed. In Section 4 and Section 5 a hardware architecture of an elliptic curve scalar multiplier is presented. This architecture uses the multiplication, squaring and inversion methods discussed in Section 3. Finally Section 6 provides concluding remarks and a summary of the research contributions documented in this report.

2. BACKGROUND

The fundamental building block for any elliptic curve-based cryptosystem is elliptic curve scalar multiplication. It is this operation that is to be performed by the co-processor. Provided in this section is an overview of the mathematics behind elliptic curve scalar multiplication, including both field arithmetic and curve arithmetic.

2.1. Arithmetic over Binary Finite Fields

The elements of the binary field GF(2^m) are interrelated through the operations of addition and multiplication. Since the additive and multiplicative inverses exist for all fields, the subtraction and division operations are also defined. Discussed in this section are basic methods for computing the sum, difference and product of two elements. Also presented is a method for computing the inverse of an element. The inverse, along with a multiplication, is used to implement division.

Addition and Subtraction: If two field elements $a, b \in$GF(2^m) are represented as polynomials $A(x) = a_{m-1}x^{m-1} + \cdots + a_1x + a_0$ and $B(x) = b_{m-1}x^{m-1} + \cdots + b_1x + b_0$ respectively, then their sum is written

$$S(x) = A(x) + B(x) = \sum_{i=0}^{m-1} (a_i + b_i)x^i. \quad (1)$$

A field of characteristic two provides two distinct advantages. First, the bit additions $a_i + b_i$ in (1) are performed modulo 2 and translate to an exclusive-OR (XOR) operation. The entire addition is computed by a component-wise XOR operation and does not require a carry chain. The second advantage is that in GF(2) the element 1 is its own additive inverse (i.e. $1 + 1 = 0$ or $1 = -1$). Hence, addition and subtraction are equivalent.

Multiplication: The product of field elements a and b is written as

$$P(x) = A(x) \times B(x) \mod F(x) = \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} a_i b_j x^{i+j} \mod F(x)$$

where $F(x)$ is the field reduction polynomial. By expanding $B(x)$ and distributing $A(x)$ through its terms we get

$$P(x) = b_{m-1} x^{m-1} A(x) + \cdots + b_1 x A(x) + b_0 A(x) \mod F(x).$$

By repeatedly grouping multiples of x and factoring out x we get

$$P(x) = (\cdots(((A(x)b_{m-1})x + A(x)b_{m-2})x + \cdots + A(x)b_1)x + A(x)b_0) \mod F(x). \tag{2}$$

A bit level algorithm can be derived from (2). However, many of the faster multiplication algorithms rely on the concept of group-level multiplication. Let g be an integer less than m and let $s = \lceil m/g \rceil$. If we define the polynomials

$$B_i(x) = \begin{cases} \displaystyle\sum_{j=0}^{g-1} b_{ig+j} x^j & 0 \le i \le s-2, \\ \displaystyle\sum_{j=0}^{(m \mod g)-1} b_{ig+j} x^j & i = s-1, \end{cases}$$

then the product of a and b is written

$$P(x) = A(x) \left(x^{(s-1)g} B_{s-1}(x) + \cdots + x^g B_1(x) + B_0(x) \right) \mod F(x).$$

In the derivation of equation (2) multiples of x were repeatedly grouped then factored out. This same grouping and factoring procedure will now be implemented for multiples of x^g arriving at

$$P(x) = (\cdots((A(x)B_{s-1}(x))x^g + A(x)B_{s-2}(x))x^g + \cdots)x^g + A(x)B_0(x) \mod F(x)$$

which can be computed using Algorithm 1.

Algorithm 1. Group-Level Multiplication

Input: $A(x)$, $B(x)$, and $F(x)$

Output: $P(x) = A(x)B(x) \mod F(x)$

$P(x) \leftarrow B_{s-1}(x)A(x) \mod F(x)$;

for $k = s - 2$ **downto** 0 **do**

$P(x) \leftarrow x^g P(x)$;

$P(x) \leftarrow B_k(x)A(x) + P(x) \mod F(x)$;

Inversion: For any element $a \in \text{GF}(2^m)$ the equality $a^{2^m-1} \equiv 1$ holds. When $a \neq 0$, dividing both sides by a results in $a^{2^m-2} \equiv a^{-1}$. Using this equality the inverse, a^{-1}, can be computed through successive field squarings and multiplications. In Algorithm 2 the inverse of an element is computed using this method.

Algorithm 2. Inversion by Square and Multiply

Input: Field element a

Output: $b \equiv a^{(-1)}$

$b \leftarrow a$;

for $i = 1$ **to** $m - 2$ **do**

$b \leftarrow b^2 * a$;

$b \leftarrow b^2$;

The primary advantage to this inversion method is the fact that it does not require hardware dedicated specifically to inversion. The field multiplier can be used to perform all required field operations.

2.2. Arithmetic over the Elliptic Curve Group

The field operations discussed in the previous section are used to perform arithmetic over an elliptic curve. This chapter is aimed at the elliptic curve defined by the non-supersingular Weierstrass equation for binary fields. This curve is defined by the equation

$$y^2 + xy = x^3 + \alpha x^2 + \beta \tag{3}$$

where the variables x and y are elements of the field GF(2^m) as are the curve parameters α and β. The points on the curve, defined by the solutions, (x, y), to (3) form an additive group when combined with the "point at infinity". This extra point is the group identity and is denoted by the symbol $\mathcal{O}$. By definition, the addition of two elements in a group results in another element of the group. As a result any point on the curve, say P, can be added to itself an arbitrary number of times and the result will also be a point on the curve. So for any integer k and point P adding P to itself $k-1$ times results in the point

$$kP = \underbrace{P + P + \cdots + P}_{k \text{ times}}.$$

Given the binary expansion $k = 2^{l-1}k_{l-1} + 2^{l-2}k_{l-2} + \cdots + 2k_1 + k_0$ the scalar multiple kP can be computed by

$$Q = kP = 2^{l-1}k_{l-1}P + 2^{l-2}k_{l-2}P + \cdots + 2k_1P + k_0P.$$

By factoring out 2, the result is

$$Q = (2^{l-2}k_{l-1}P + 2^{l-3}k_{l-2}P + \cdots + k_1P)2 + k_0P.$$

By repeating this operation it is seen that

$$Q = (\cdots((k_{l-1}P)2 + k_{l-2}P)2 + \cdots + k_1P)2 + k_0P$$

which can be computed by the well known (left-to-right) double and add method for scalar multiplication shown in Algorithm 3.

Two basic operations required for elliptic curve scalar multiplication are point `ADD` and point `DOUBLE`. The mathematical definitions for these operations are derived from the curve equation in (3). Consider the points P_1 and P_2 represented by the coordinate pairs (x_1, y_1) and (x_2, y_2) respectively. Then the coordinates, (x_a, y_a), of point $P_a = P_1 + P_2$ (or `ADD`(P_1, P_2)) are computed using the equations

$$x_a = \left(\frac{y_1 + y_2}{x_1 + x_2}\right)^2 + \frac{y_1 + y_2}{x_1 + x_2} + x_1 + x_2 + \alpha$$
$$y_a = \left(\frac{y_1 + y_2}{x_1 + x_2}\right)(x_1 + x_a) + x_a + y_1.$$

Similarly the coordinates (x_d, y_d) of point $P_d = 2P_1$ (or `DOUBLE`(P_1)) are computed using the equations

$$x_d = x_1^2 + \left(\frac{\beta}{x_1^2}\right)$$
$$y_d = x_1^2 + \left(x_1 + \frac{y_1}{x_1}\right)x_d + x_d.$$

Algorithm 3. Scalar Multiplication by Double and Add Method

Input: Integer $k = (k_{l-1}, k_{l-2}, \ldots, k_1, k_0)_2$, Point P

Output: Point $Q = kP$

$Q \leftarrow \mathcal{O}$;

if $(k_{l-1} == 1)$ **then**

$Q \leftarrow P$;

for $i = l - 2$ **downto** 0 **do**

$Q \leftarrow \texttt{DOUBLE}(Q)$;

if $(k_i == 1)$ **then**

$Q \leftarrow \texttt{ADD}(Q, P)$;

So the addition of two points can be computed using two field multiplications, one field squaring, eight field additions and one field inversion. The double of a point can be computed using two field multiplications, one field squaring, six field additions and one field inversion.

3. HIGH PERFORMANCE FINITE FIELD ARITHMETIC

In order to optimize the curve arithmetic discussed in Section 2.2 the underlying field operations must be implemented in a fast and efficient way. The required field arithmetic operations are addition, multiplication, squaring and inversion. Each of these operations have been implemented in hardware for use in the prototype discussed in Section 5. Generally speaking, field multiplication has the greatest effect on the performance of the entire elliptic curve scalar multiplication.[1] For this reason, focus will be primarily on the field multiplier when discussing hardware architectures for field arithmetic.

This section is organized as follows. Section 3.1 presents a hardware architecture designed to perform finite field multiplication. In Section 3.2 the ideas presented for multiplication are extended to create a hardware architecture optimized for squaring. Section 3.3 gives a method for inversion due to Itoh and Tsujii. This method does not require any additional hardware but instead uses the multiplication and squaring units described in Sections 3.1 and 3.2. Section 3.4 gives a description of a comparator/adder

[1] Inversion takes much longer than multiplication, but its effect on performance can be greatly reduced through use of projective coordinates. This is discussed in greater detail in Section 4.1.

which both compares and adds finite field elements. Finally, Section 3.5 summarizes results gleaned from a hardware prototype of each arithmetic unit/routine.

3.1. Multiplication

In [11] a digit serial multiplier is proposed which is based on look-up tables. This method was implemented in software for the field GF(2^{163}) and reported in [14]. To the best of our knowledge this performance of 0.540 μ-seconds for a single field multiplication is the fastest reported result for a software implementation. In this section the possibilities of using this look-up table-based algorithm in hardware will be explored.

First to be described in this section is the algorithm used for multiplication. Then we present a hardware structure designed to compute $R(x)W(x) \mod F(x)$ where $R(x)$ and $W(x)$ are polynomials with degrees $g-1$ and $m-1$ respectively and $g << m$. A description of the multiplier's data path follows. In conclusion there will be a discussion behind the reasons for the choice of digit sizes.

Multiplication Algorithm: The computations of

$$P(x) \leftarrow x^g P(x) \mod F(x) \text{ and}$$
$$P(x) \leftarrow B_k(x)A(x) + P(x) \mod F(x)$$

from the **for** loop of Algorithm 1 on page 7 can be broken up into the following steps.

$$V_1 = x^g \sum_{i=0}^{m-g-1} p_i x^i,$$
$$V_2 = x^g \sum_{i=m-g}^{m-1} p_i x^i \mod F(x)$$
$$V_3 = B_k(x)A(x) \mod F(x) \text{ and}$$
$$P(x) = V_1 + V_2 + V_3$$

Note that V_1 is a g-bit shift of the lower $m-g$ bits of $P(x)$. V_2 is a g-bit shift of the upper g bits of $P(x)$ followed by a modular reduction. V_3 requires a polynomial multiplication and reduction where the operand polynomials have degree $g-1$ and $m-1$. Algorithm 1 can be modified to create Algorithm 4.

In [11] polynomials V_2 and V_3 are computed with the assistance of look-up tables mainly for software implementation. The look-up tables used to compute V_2 and V_3 are referred to as the M-Table and T-Table respectively. The M-Table is addressed by the bit string $(p_{m-1}, p_{m-2}, \ldots, p_{m-g})$ interpreted as the integer $2^{g-1}p_{m-1} + 2^{g-2}p_{m-2} + \cdots + p_{m-g}$. Similarly the T-Table is addressed by the coefficients of $B_k(x)$, or the integer $B_k(x = 2)$. The elements of the M-Table are a function of the reduction polynomial $F(x)$ and can be precomputed. The elements of the T-Table are a function

Algorithm 4. Efficient Group Level Multiplication

Input: $A(x)$, $B(x)$, and $F(x)$

Output: $P(x) = A(x)B(x) \mod F(x)$

$P(x) \leftarrow B_{s-1}(x)A(x) \mod F(x)$;

for $k = s - 2$ **downto** 0 **do**

$V_1 \leftarrow x^g \sum_{i=0}^{m-g-1} p_i x^i$;

$V_2 \leftarrow x^g \sum_{i=m-g}^{m-1} p_i x^i \mod F(x)$;

$V_3 \leftarrow B_k(x)A(x) \mod F(x)$;

$P(x) \leftarrow V_1 + V_2 + V_3$;

of $A(x)$ and hence are dynamic. These values need to be computed each time a new $A(x)$ is used.

Computation of $R(x)W(x) \mod F(x)$: Instead of using tables, below the polynomials V_2 and V_3 are computed on the fly. The computation of V_2 and V_3 are similar in that they both require a multiplication of two polynomials followed by a reduction, where the first polynomial has degree $g - 1$ and the other has degree less than m. This is obvious for V_3 and can be shown easily for V_2. Note that

$$\begin{aligned} V_2 &= p_{m-1}x^{m+g-1} + \cdots + p_{m-g+1}x^{m+1} + p_{m-g}x^m \mod F(x) \\ &= x^m \left(p_{m-1}x^{g-1} + \cdots + p_{m-g+1}x + p_{m-g}\right) \mod F(x). \end{aligned}$$

The field reduction polynomial $F(x) = x^m + x^d + \cdots + 1$ provides us the equality $x^m \equiv x^d + \cdots + 1$. Substituting for x^m we see that

$$V_2 = \left(x^d + \cdots + 1\right)\left(p_{m-1}x^{g-1} + \cdots + p_{m-g+1}x + p_{m-g}\right) \mod F(x).$$

Provided $d + g < m$, V_2 results in a polynomial of degree less than m which does not need to be reduced. Since d is relatively small for all five NIST polynomials, it is reasonable to assume that $d+g < m$. For the remainder of this chapter, this assumption is used.

With this said, the following method can be used to compute both V_2 and V_3. Consider the polynomial multiplication and reduction $R(x)W(x) \mod F(x)$ where

$R(x) = \sum_{i=0}^{g-1} r_i x^i$ and $W(x)$ is a polynomial with degree less than m. Then

$$\begin{aligned} R(x)W(x) \mod F(x) = & r_{g-1}(x^{g-1}W(x) \mod F(x)) \\ & + r_{g-2}(x^{g-2}W(x) \mod F(x)) \\ & \vdots \\ & + r_1(xW(x) \mod F(x)) \\ & + r_0(W(x) \mod F(x)) \end{aligned}$$

The value $x^i W(x) \mod F(x)$ is just a shifted and reduced version of $x^{i-1}W(x) \mod F(x)$. So each value $x^i W(x) \mod F(x)$ can be generated sequentially starting with $x^0 W(x)$ as shown in Figure 1. When using a reduction polynomial with a low Hamming weight, such as a trinomial or pentanomial, these terms can be computed quickly at very little cost. Once these values are determined, the final result is computed using a g-input modulo 2 adder. The inputs to the adder are enabled by their corresponding coefficient r_i. This is shown in Figure 2. Note that the polynomial $x^i W(x)$ affects the output of the adder only if the coefficient bit r_i is a one. Otherwise the input associated with $x^i W(x)$ is driven with zeros.

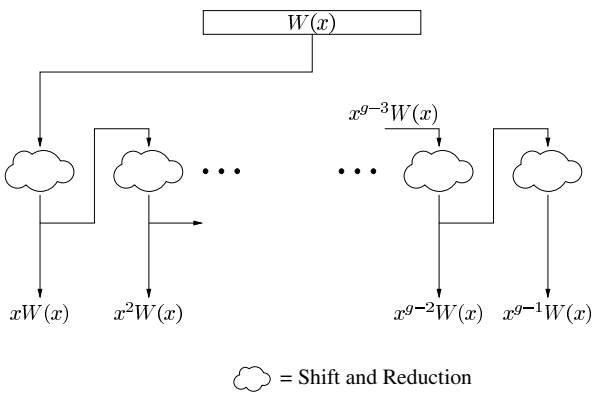

Figure 1. Generating $x^i W(x) \mod F(x)$

Each individual output bit of the g-operand mod 2 adder is computed using $g-1$ XOR gates and g AND gates. The AND gates are used to enable each input bit and the XOR gates compute the mod 2 addition. Figure 3 demonstrates how this is done. The depth of the logic in the figure is linearly related to g.

This method for multiplication is implemented for computation of both V_2 and V_3. In the case of V_3, the polynomial $W(x)$ has degree $m-1$ and will change for every

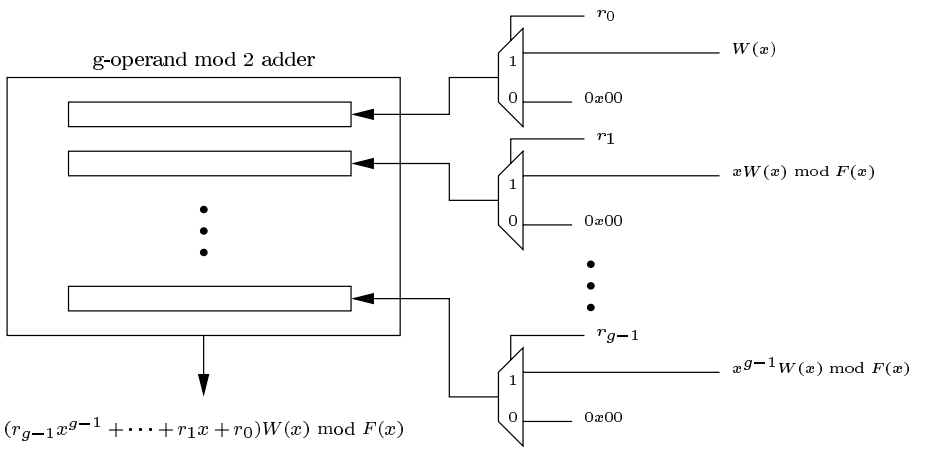

Figure 2. Computing $R(x)W(x) \mod F(x)$

field multiplication. For V_2 the polynomial $W(x)$ has degree d and is fixed. The value d is the degree of the second leading non-zero coefficient of $F(x)$. For reasonable digit sizes this computation can be performed in a single clock cycle.

Multiplier Data Path: The multiplier's data path connecting the V_2 and V_3 generators along with the adder used to compute $P(x) = V_1 + V_2 + V_3$ is shown in Figure 4. A buffer is inserted at the output of the V_3 generator to separate its delay from the delay of the adder for $V_1 + V_2 + V_3$. This, in effect, increases the maximum possible value for the digit size g. If added by itself, this buffer would add a cycle of latency to the multiplier's performance time. This extra cycle is compensated for by bypassing the $P(x)$ register and driving the multiplier's output with the output of the 3-operand mod2 adder. It is important to note that the delay of the 3-operand mod2 adder is being merged with the delay of the bus which connects the multiplier to the rest of the design. In this case the relatively relaxed bus timing has room to accommodate the delay.

Choice of Digit Size: The multiplier will complete a multiplication in $\lceil m/g \rceil$ clock cycles. Since this is a discrete value, the performance may not change for every value of g. To minimize cost of the multiplier (which increases with g) the smallest digit size g should be chosen for a given performance $\lceil m/g \rceil$. For example, the digit sizes $g = 21$ and $g = 22$ for field size $m = 163$ result in the same performance, $\lceil \frac{163}{21} \rceil = \lceil \frac{163}{22} \rceil = 8$, but $g = 22$ requires a larger multiplier.

Implementation results of a prototype of this multiplier for the field GF(2^{163}) and NIST polynomial for various digit sizes are shown in Table 2. For each digit size, the table lists the corresponding cycle performance and resource cost. A maximum digit

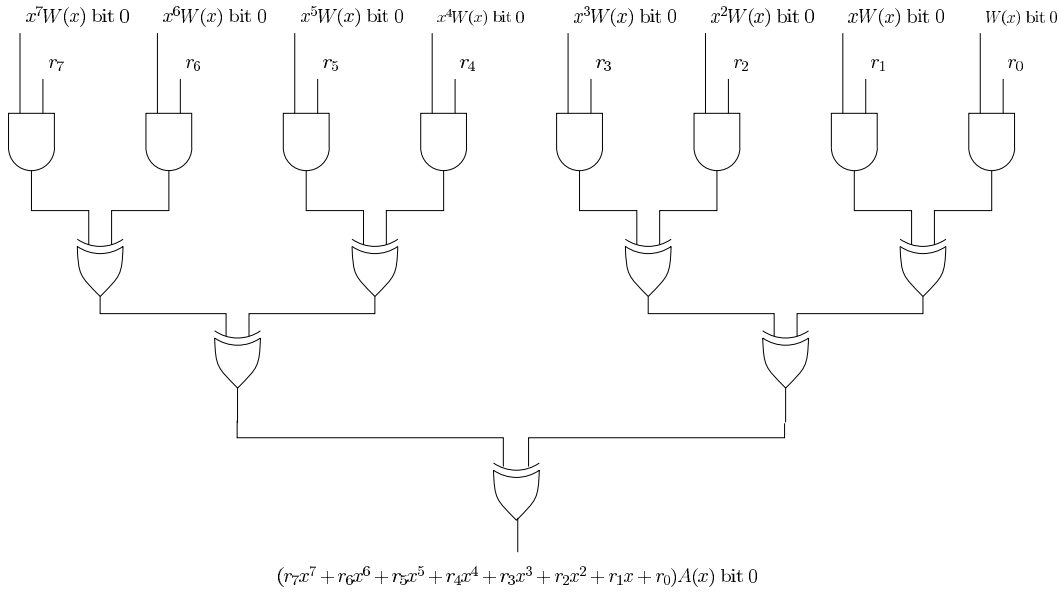

Figure 3. Computation of a Single Bit in $R(x)W(x) \mod F(x)$

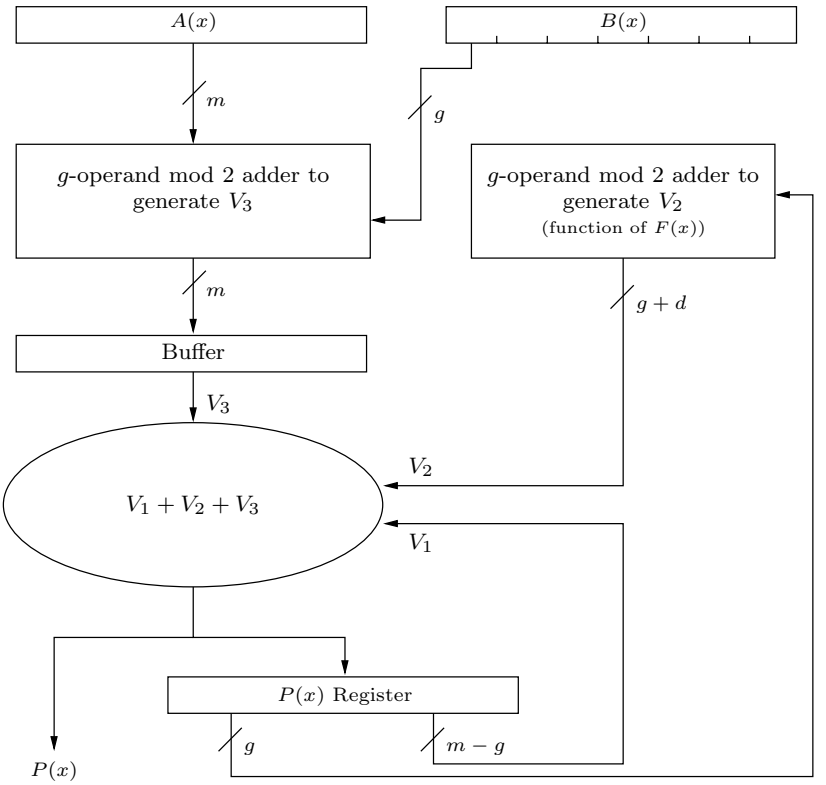

Figure 4. Multiplier Data-Path

size of $g = 41$ is a good choice for several reasons. First, as the performance cost of the actual field multiplication decreases, the relative cost of loading and unloading the multiplier increases. So as the digit size increases, its affect on the total performance (including time to load and unload the multiplier) decreases. Second, results showed that $g > 41$ had difficulty meeting timing at the target operating frequency of 66 MHz.

3.2. Squaring

While squaring is a specific case of general multiplication and can be performed by the multiplier, performance can be improved significantly by optimizing the architecture specifically for the case of squaring. The square of an element a represented by $A(x)$ involves two mathematical steps. The first is the polynomial multiplication of $A(x)$ resulting in

$$A^2(x) = a_{m-1}x^{2m-2} + \cdots + a_2x^4 + a_1x^2 + a_0.$$

Table 2. Performance/Cost Trade-off for Multiplication over GF(2^{163})

Digit Size	Performance in clock cycles	# LUTs	# Flip Flops
$g = 1$	163	677	670
$g = 4$	41	854	670
$g = 28$	6	3,548	670
$g = 33$	5	4,040	670
$g = 41$	4	4,728	670

The second is the reduction of this polynomial modulo $F(x)$. Assuming that m is an odd integer, which is the case for all five NIST recommended binary fields, if the terms with degree greater than $m-1$ are separated and x^{m+1} is factored out where possible the result will be $A^2(x) = A_h(x)x^{m+1} + A_l(x)$ where

$$A_h(x) = a_{m-1}x^{m-3} + \cdots + a_{(\frac{m+3}{2})}x^2 + a_{(\frac{m+1}{2})}$$
$$A_l(x) = a_{(\frac{m-1}{2})}x^{m-1} + \cdots + a_1x^2 + a_0,$$

The polynomial $A_l(x)$ has degree less than m and does not need to be reduced. The product $A_h(x)x^{m+1}$ may have degree as large as $2m-2$. The reduction polynomial gives us the equality $x^m = x^d + \cdots + 1$. Multiplying both sides by x, we get $x^{m+1} = x^{d+1} + \cdots + x$. So

$$A_h(x)x^{m+1} = A_h(x)\left(x^{d+1} + \cdots + x\right).$$

This multiplication can be performed using a method similar to the one described in Section 3.1. The same architecture used to compute $R(x)W(x) \mod F(x)$ in the multiplier is used here to compute $x^{m+1}A_h(x)$. The digit size is set to $g = d + 2$ and the elements of g-operand mod 2 adder are generated from $A_h(x)$. $A_h(x)$ is in turn generated by expanding $A(x)$ (i.e., inserting zeros between the coefficient bits of $A(x)$). Since the digit size is set to $d + 2$, the multiplication is completed in a single cycle. This method only works if $d + 2 < m$ which is the case for each of the NIST polynomials. Figure 5 shows the data flow for the squaring operation. Note that the flow does not include any buffers and so is implemented in pure combinational logic.

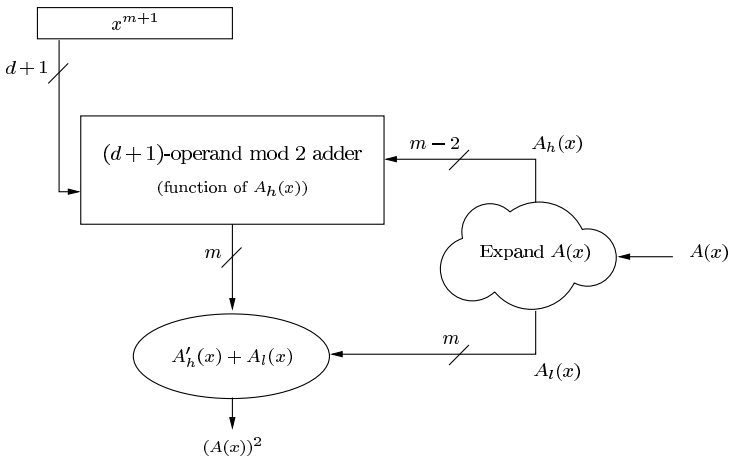

Figure 5. Data-Path of the Squaring Unit

The prototype of this squaring unit for field GF(2^{163}) using the NIST reduction polynomial runs at 66 MHz and is capable of performing a squaring operation in a single clock cycle. This implementation requires 330 LUTs and 328 Flip Flops.

3.3. Inversion

The inversion method described in Algorithm 2 on page 7 requires $m-1$ squarings and $m-2$ multiplications. In order to accurately estimate the cycle performance of the inversion, consideration must be given to the performance of the multiplication and squaring units as well as the time required to load and unload these units. The architecture of the elliptic curve scalar multiplier will be discussed in detail in Section 5. For now, it is sufficient to know that the arithmetic units are loaded using two independent m bit data buses and unloaded using a single m bit data bus. The operands are stored in a dual port memory which takes two clock cycles to read from and one cycle to write to. These combined makes three cycles that are required to both load and unload any arithmetic unit. Further analysis assumes that these three cycles remain constant for all m. If C_s and C_m denote the number of clock cycles required to complete a squaring and multiplication respectively, then an inversion can be completed in

$$(C_s + 3)(m-1) + (C_m + 3)(m-2)$$

clock cycles. For the field GF(2^{163}) where $C_s = 1$ and $C_m = 4$, this translates to 1775 clock cycles.

Performance can be improved by using Algorithm 5 due to Itoh and Tsujii [13]. This algorithm is derived from the equation $a^{(-1)} \equiv a^{2^m - 2} \equiv \left(2^{2^m - 1} - 1\right)^2$

which is true for any non-zero element $a \in$GF(2^m). From

$$a^{2^t-1} \equiv \begin{cases} \left(a^{2^{t/2}-1}\right)^{2^{t/2}} \left(a^{2^{t/2}-1}\right) & \text{for } t \text{ even}, \\ a\left(a^{2^{t-1}-1}\right)^2 & \text{for } t \text{ odd}, \end{cases} \tag{4}$$

the computation required for the exponentiation $2^{2^{m-1}-1}$ can be iteratively broken down. Algorithm 5 requires $\lfloor \log_2(m-1) \rfloor + H(m-1) - 1$ multiplications and $m-1$ squarings. Using the notation defined earlier, this translates to

$$(C_s + 3)(m-1) + (C_m + 3)(\lfloor \log_2(m-1) \rfloor + H(m-1) - 1)$$

clock cycles. For GF(2^{163}) this translates to 711 clock cycles.

Algorithm 5. Optimized Inversion by Square and Multiply

Inputs: Field element $a \neq 0$,

Binary representation of $m - 1 = (m_{l-1}, \ldots, m_2, m_0)_2$

Output: $b \equiv a^{(-1)}$

$b \leftarrow a^{m_{l-1}}$;

$e \leftarrow 1$;

for $i = l - 2$ **downto** 0 **do**

  $b \leftarrow b^{2^e} b$;

  $e \leftarrow 2e$;

  if $(m_i == 1)$ **then**

    $b \leftarrow b^2 a$;

    $e = e + 1$;

$b \leftarrow b^2$;

Now, the majority of the time spent for each squaring operation is used to load and unload the squaring unit (three out of the four cycles). Algorithm 5 requires several sequences of repetitive squaring (i.e. computations of the form x^{2^t}). These repeated squarings do not require intermediate values to be stored outside the squaring unit. By modifying the squaring unit to support the *re-square* of an element, most of the memory accesses otherwise required to load and unload the squaring unit are eliminated. In fact,

the squaring unit only needs to be loaded and unloaded once for each multiplication. Hence the number of clock cycles is reduced to

$$(C_s(m-1) + 3(\lfloor \log_2(m-1) \rfloor + H(m-1) - 1)) \\ + (C_m + 3)(\lfloor \log_2(m-1) \rfloor + H(m-1) - 1)$$

clock cycles. For the field GF(2^{163}) with $C_s = 1$ and $C_m = 4$, this results in 252 clock cycles.

This is a competitive value since a typical hardware implementation of the Extended Euclidean Algorithm (EEA) is expected to complete an inversion in approximately $2m$ clock cycles or 326 cycles for GF(2^{163}). This corresponds to a 60 clock cycle reduction or 20% performance improvement without requiring hardware dedicated specifically for inversion. Table 3 lists the performance numbers of the previously mentioned inversion methods when implemented over the field GF(2^{163}).

Table 3. Comparison of Various Inversion Methods for GF(2^{163})

Method	# Squarings	# Multiplications	# Cycles
Square & Multiply	$m-1$	$m-2$	1127
Itoh & Tsujii	$m-1$	$\lfloor \log_2(m-1) \rfloor + H(m) - 1$	711
Itoh & Tsujii w/ *re-square*	$m-1$	$\lfloor \log_2(m-1) \rfloor + H(m) - 1$	252
EEA	-	-	326

The actual time to complete an inversion using the ECC co-processor architecture discussed in Section 5 is 259 clock cycles. The 7 extra cycles are due to control related instructions executed in the micro-sequencer.

3.4. Comparator/Adder

The primary purpose of the Comparator/Adder is to compute the sum of two field elements. This is done with an array of m exclusive OR gates. To minimize register usage as well as time to complete the addition, the sum of the two operands is the only value stored in a register. In this way, the sum is available immediately after the operands are loaded into the Comparator/Adder. In other words, it takes no extra clock cycles to complete a finite field addition.

In addition to computing the sum of two finite field elements, the Comparator/Adder also acts as a comparator. The comparison is performed by taking the logical NOR of all the bits in the sum register. If the result is a one, then the sum is zero and the two operands are equal. If operand a is set to zero, then operand b can be tested for zero.

The logic depth for the zero detect circuitry (the m-bit NOR gate) is $\log_2(m)$ and is registered before being sent out of the module. Figure 6 provides a functional diagram of the Comparator/Adder.

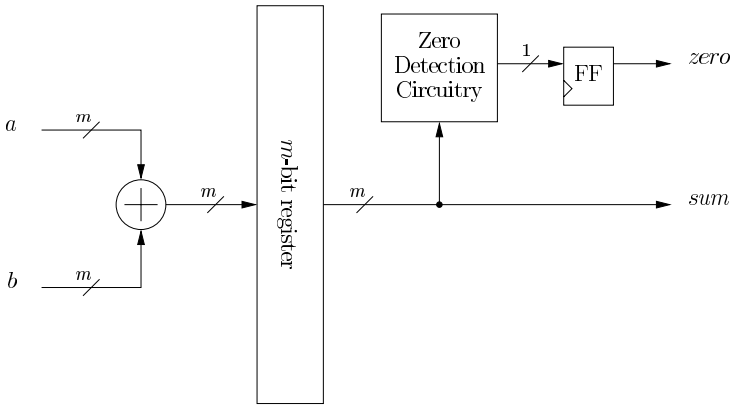

Figure 6. Data-Path of the Comparator/Adder

3.5. Remarks

In this section, we have discussed hardware architectures designed to perform finite field addition, multiplication and squaring. Also discussed was an efficient method for inversion which uses the squaring and multiplication units. The performance results associated with these arithmetic units are summarized in Table 4.

4. ECC SCALAR MULTIPLICATION

The section is organized as follows. Section 4.1 introduces projective coordinates and discusses some of the reasons for using a projective system. Section 4.2 presents two methods for recoding the scalar. They are non-adjacent form (NAF) and τ-adic non-adjacent form (τ-NAF).

4.1. Choice of Coordinate Systems

Projective coordinates allow the inversion required by each `DOUBLE` and `ADD` to be eliminated at the expense of a few extra field multiplications. The benefit is measured by the ratio of the time to complete an inversion to the time to complete a multiplication. The inversion algorithm proposed by Itoh and Tsujii [13] will be used

Table 4. Performance of Finite Field Operations

Operation ($g = 41$)	# Cycles	# Cycles Including Initial and Final Data Movement
Multiplication	4	7
Squaring	1	4
Addition	0	3
Inversion	256	259

and therefore, the above ratio is guaranteed to be larger than $\lfloor \log_2(m-1) \rfloor$ and could be larger depending on the efficiency of the squaring operations. Therefore, projective coordinates will provide us the best performance for NIST curves. Several flavors of projective coordinates have been proposed over the last few years. The prominent ones are *Standard* [21], *Jacobian* [4, 12] and López & Dahab [18] projective coordinates.

If the affine representation of P be denoted as (x, y) and the projective representation of P be denoted as (X, Y, Z), then the relation between affine and projective coordinates for the Standard system is

$$x = \frac{X}{Z} \quad \text{and} \quad y = \frac{Y}{Z}.$$

For Jacobian projective coordinates the relation is

$$x = \frac{X}{Z^2} \quad \text{and} \quad y = \frac{Y}{Z^3}.$$

Finally for López & Dahab's, the relation between affine and projective coordinates is

$$x = \frac{X}{Z} \quad \text{and} \quad y = \frac{Y}{Z^2}.$$

For López & Dahab's system the projective equation of the elliptic curve in (3) then becomes

$$Y^2 + XYZ = X^3Z + \alpha X^2Z^2 + \beta Z^4.$$

It is important to note that when using the left-to-right double and add method for scalar multiplication all point additions are of the form $\texttt{ADD}(P, Q)$. The base point P is never modified and as a result will maintain its affine representation (i.e. $P = (x, y, 1)$). The constant Z coordinate significantly reduces the cost of point addition (from 14 field multiplications down to 10). The addition of two distinct points $(X_1, Y_1, Z_1) + (X_2, Y_2, 1) = (X_a, Y_a, Z_a)$ using *mixed* coordinates (one projective point and one

affine point) is then computed by

$$\begin{array}{ll} A = Y_2 \cdot Z_1^2 + Y_1 & E = A \cdot C \\ B = X_2 \cdot Z_1 + X_1 & X_a = A^2 + D + E \\ C = Z_1 \cdot B & F = X_a + X_2 \cdot Z_a \\ D = B^2 \cdot (C + \alpha \cdot Z_1^2) & G = X_a + Y_2 \cdot Z_a \\ Z_a = C^2 & Y_a = E \cdot F + Z_a \cdot G \end{array} \tag{5}$$

Similarly, the double of a point (X_1, Y_1, Z_1) is $(X_d, Y_d, Z_d) = 2(X_1, Y_1, Z_1)$ is computed by

$$\begin{array}{l} Z_d = Z_1^2 \cdot X_1^2 \\ X_d = X_1^4 + \beta \cdot Z_1^4 \\ Y_d = \beta \cdot Z_1^4 \cdot Z_d + X_d \cdot (\alpha \cdot Z_d + Y_1^2 + \beta \cdot Z_1^4) \end{array} \tag{6}$$

In Table 5, the number of field operations required for the affine, Standard, Jacobean and López & Dahab coordinate systems are provided. In the table the symbols $\mathcal{M}$, $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{I}$ denote field multiplication, squaring, addition and inversion respectively.

Table 5. Comparison of Projective Point Systems

System	Point Addition	Point Doubling
Affine	$2\mathcal{M} + 1\mathcal{S} + 8\mathcal{A} + 1\mathcal{I}$	$3\mathcal{M} + 2\mathcal{S} + 4\mathcal{A} + 1\mathcal{I}$
Standard	$13\mathcal{M} + 1\mathcal{S} + 7\mathcal{A}$	$7\mathcal{M} + 5\mathcal{S} + 4\mathcal{A}$
Jacobian	$11\mathcal{M} + 4\mathcal{S} + 7\mathcal{A}$	$5\mathcal{M} + 5\mathcal{S} + 4\mathcal{A}$
López & Dahab	$10\mathcal{M} + 4\mathcal{S} + 8\mathcal{A}$	$5\mathcal{M} + 5\mathcal{S} + 4\mathcal{A}$

The projective coordinate system defined by López and Dahab will be used since it offers the best performance for both point addition and point doubling.

4.2. Scalar Multiplication using Recoded Integers

The binary expansion of an integer k is written as $k = \sum_{i=0}^{l-1} k_i 2^i$ where $k_i \in \{0, 1\}$. For the case of elliptic curve scalar multiplication the length l is approximately equal to m, the degree of the extension field. Assuming an average Hamming weight, a scalar multiplication will require approximately $l/2$ point additions and $l - 1$ point

doubles. Several recoding methods have been proposed which in effect reduce the number of additions. In this section two methods are discussed, namely NAF [9, 29] and τ-adic NAF [16, 29].

Scalar Multiplication using Binary NAF: The symbols in the binary expansion are selected from the set $\{0, 1\}$. If this set is increased to $\{0, 1, -1\}$ the expansion is referred to as *signed binary* (SB) representation. When using this representation, the double and add scalar multiplication method must be slightly modified to handle the -1 symbol (often denoted as $\bar{1}$). If the expansion $k'_{l-1}2^{l-1} + \cdots + k'_1 2 + k'_0$ where $k'_i \in \{0, 1, \bar{1}\}$ is denoted by $(k'_{l-1}, \ldots, k'_1, k'_0)_{SB}$, then Algorithm 6 computes the scalar multiple of point P. The negative of the point (x, y) is $(x, x + y)$ and can be computed

Algorithm 6. Scalar Multiplication for Signed Binary Representation

Input: Integer $k = (k'_{l-1}, k'_{l-2}, \ldots, k'_1, k'_0)_{\text{SB}}$, Point P

Output: Point $Q = kP$

$Q \leftarrow \mathcal{O}$;

if $(k'_{l-1} \neq 0)$ **then**

$Q \leftarrow k'_{l-1}P$;

for $i = l - 2$ **downto** 0 **do**

$Q \leftarrow \texttt{DOUBLE}(Q)$;

if $(k'_i \neq 0)$ **then**

$Q \leftarrow \texttt{ADD}(Q, k'_i P)$;

with a single field addition. The signed binary representation is redundant in the sense that any given integer has more than one possible representation. For example, 17 can be represented by $(1001)_{\text{SB}}$ as well as $(101\bar{1})_{\text{SB}}$.

Interest here is in a particular form of this signed binary representation called NAF or non-adjacent form. A signed binary integer is said to be in NAF if there are no adjacent non-zero symbols. The NAF of an integer is unique and it is guaranteed to be no more than one symbol longer than the corresponding binary expansion. The primary advantage gained from NAF is its reduced number of non-zero symbols. The average Hamming weight of a NAF is approximately $l/3$ [29] compared to that of the binary expansion which is $l/2$. As a result, the running time of elliptic curve scalar multiplication when using binary NAF is reduced to $(l + 1)/3$ point additions and l point doubles. This represents a significant reduction in run time.

In [29], Solinas provides a straightforward method for computing the NAF of an integer. This method is given here in Algorithm 7.

Algorithm 7. Generation of Binary NAF

Input: Positive integer k

Output: $k' = \text{NAF}(k)$

$i \leftarrow 0$;

while $(k > 0)$ **do**

  if $(k \equiv 1 \pmod 2)$ **then**

    $k'_i \leftarrow 2 - (k \mod 4)$;

    $k \leftarrow k - k'_i$;

  else

    $k'_i \leftarrow 0$;

  $k \leftarrow k/2$;

  $i \leftarrow i + 1$;

Scalar Multiplication using τ-NAF: Anomalous Binary Curves (ABC's), first proposed for cryptographic use in [16], provide an efficient implementation when the scalar is represented as a complex algebraic number. ABC's, often referred to as the Koblitz curves, are defined by

$$y^2 + xy = x^3 + \alpha x^2 + 1 \tag{7}$$

with $\alpha = 0$ or $\alpha = 1$. The advantage provided by the Koblitz curves is that the `DOUBLE` operation in Algorithm 6 can be replaced with a second operation, namely Frobenius mapping, which is easier to perform.

If point (x, y) is on a Koblitz curve then it can be easily checked that (x^2, y^2) is also on the same curve. Moreover, these two points are related by the following Frobenius mapping

$$\tau(x, y) = (x^2, y^2)$$

where τ satisfies the quadratic equation

$$\tau^2 + 2 = \mu\tau. \tag{8}$$

In (8), $\mu = (-1)^{1-\alpha}$ and α is the curve parameter in (7) and is 0 or 1 for the Koblitz curves.

The integer k can be represented with radix τ using signed representation. In this case, the expansion is written

$$k = \kappa_{l-1}\tau^{l-1} + \cdots \kappa_1\tau + \kappa_0,$$

where $\kappa_i \in \{0, 1, \bar{1}\}$. Using this representation, Algorithm 6 can be rewritten, replacing the DOUBLE(Q) operation with τQ or a Frobenius mapping of Q. The modified algorithm is shown in Algorithm 8. Since τQ is computed by squaring the coordinates of Q, this suggests a possible speed up over the DOUBLE and ADD method.

Algorithm 8. Scalar Multiplication for τ-adic Integers

Input: Integer $k = (\kappa_{l-1}, \kappa_{l-2}, \ldots, \kappa_1, \kappa_0)_\tau$, Point P

Output: Point $Q = kP$

$Q \leftarrow \mathcal{O}$;

if $(\kappa_{l-1} \neq 0)$ **then**

$Q \leftarrow \kappa_{l-1}P$;

for $i = l - 2$ **downto** 0 **do**

$Q \leftarrow \tau Q$;

if $(\kappa_i \neq 0)$ **then**

$Q \leftarrow$ ADD$(Q, \kappa_i P)$;

This complex representation of the integer can be improved further by computing its non-adjacent form. Solinas proved the existence of such a representation in [29] by providing an algorithm which computes the τ-adic non-adjacent form or τ-NAF of an integer. This algorithm is provided here in Algorithm 9. In most cases, the input to Algorithm 9 will be a binary integer, say k (i.e. $r_0 = k$ and $r_1 = 0$). If k has length l then TNAF(k) will have length $2l$, roughly twice the length of NAF(k).

The length of the representation generated by Algorithm 9 can be reduced by either preprocessing the integer k, as is done in [29], or by post processing the result. A method for post processing the output of Algorithm 9 is presented here.

Remember that $\tau(x, y) = (x^2, y^2)$. Since $z^{2^m} = z$ for all $z \in$GF(2^m), it follows that

$$\tau^m(x, y) = (x^{2^m}, y^{2^m}) = (x, y).$$

This relation gives us the general equality

$$(\tau^m - 1)P \equiv 0$$

Algorithm 9. Generation of τ-adic NAF

Input: $r_0 + r_1\tau$ where $r_0, r_1 \in \mathbb{Z}$

Output: u =TNAF($r_0 + r_1\tau$)

$i \leftarrow 0$;

while ($r_0 \neq 0$ or $r_1 \neq 0$) **do**

  if ($r_0 \equiv 1 \pmod 2$) **then**

    $u_i \leftarrow 2 - (r_0 - 2r_1 \mod 4)$;

    $r_0 \leftarrow r_0 - u_i$;

  else

    $u_i \leftarrow 0$;

  $t \leftarrow r_0$;

  $r_0 \leftarrow r_1 + \mu r_0/2$;

  $r_1 \leftarrow -t/2$;

  $i \leftarrow i + 1$;

where P is a point on a Koblitz curve. As a result, any integer k expressed with radix τ can be reduced modulo $\tau^m - 1$ without changing the scalar multiple kP. This reduction is performed easily with a few polynomial additions. Consider the τ-adic integer

$$u = u_{2m-1}\tau^{2m-1} + \cdots + u_{m+1}\tau^{m+1} + u_m\tau^m + u_{m-1}\tau^{m-1} + \cdots + u_1\tau + u_0.$$

Factoring out τ^m wherever possible, the result is

$$\begin{aligned} u = (u_{2m-1}\tau^{m-1} + \cdots + u_{m+1}\tau + u_m)\tau^m \\ +(u_{m-1}\tau^{m-1} + \cdots + u_1\tau + u_0) \end{aligned}$$

Substituting τ^m with 1 and combining terms results in

$$u = ((u_{2m-1} + u_{m-1})\tau^{m-1} + \cdots + (u_{m+1} + u_1)\tau + (u_m + u_0).$$

The output of Algorithm 9 is approximately twice the length of the input but may be slightly larger. Assuming the length of the input to be approximately m symbols, the reduction method must be capable of reducing τ-adic integers with length slightly greater $2m$. Algorithm 10 describes this method for reduction.

Algorithm 10. Reduction mod τ^m

Input: $u = u_{l-1}\tau^{l-1} + \cdots + u_1\tau + u_0$ with $m \leq l < 3m$

Output: v =REDUCE_TM(u)

$v \leftarrow 0$;

if $(l > 2m)$ **then**

$v \leftarrow (u_{l-1}\tau^{l-2m-1} + \cdots + u_{2m+1}\tau + u_{2m})$;

if $(l > m)$ **then**

$v \leftarrow v + (u_{2m-1}\tau^{m-1} + \cdots + u_{m+1}\tau + u_m)$;

$v \leftarrow v + (u_{m-1}\tau^{m-1} + \cdots + u_1\tau + u_0)$;

Now the result of Algorithm 10 has length m but is no longer in τ-adic NAF form. There may be adjacent non-zero symbols and the symbols are not restricted to the set $\{0, 1, \bar{1}\}$.

The input of Algorithm 9 is of the form $r_0 + r_1\tau$ where $r_0, r_1 \in \mathbb{Z}$. The output is the τ-adic representation of the input. For $v \in \mathbb{Z}[\tau]$ we can write

$$\begin{aligned} v &= v_{m-1}\tau^{m-1} + \cdots + v_2\tau^2 + v_1\tau + v_0 \\ &= v_{m-1}\tau^{m-1} + \cdots + v_2\tau^2 + \text{TNAF}(v_1\tau + v_0) \end{aligned}$$

Now the two least significant symbols of v are in τ-adic NAF. Repeating this procedure for every bit in v the entire string can be converted to τ-adic NAF. This process is described in Algorithm 11.

The output of Algorithm 11 is in τ-adic NAF and has a length of approximately m symbols. If the result is larger than m symbols, it is possible to repeat Algorithms 10 and 11 to further reduce the length. Algorithms 9, 10 and 11 have been implemented in C and were used to generate test vectors for the prototype discussed later in this section. During testing, it was found that a single pass of these algorithms generates a τ-adic representation with average length of m and a maximum length of $m + 5$.

Like radix 2 NAF the τ-adic NAF uses the symbol set $\{1, 0, \bar{1}\}$ and has an average Hamming weight of approximately $l/3$ for an l-bit integer [29]. So Algorithm 8 has a running time of $l/3$ point additions and $l - 1$ Frobenius mappings.

Summary and Analysis: A point addition using López & Dahab's projective coordinates requires ten field multiplications, four field squarings and eight field additions. A point double requires five field multiplications, five field squarings and four field additions. Using this information, the run time for scalar multiplication can be written in terms of field operations. Typically scalar multiplication is measured in terms of field

Algorithm 11. Regeneration of τ-adic NAF

Input: $v = v_{m-1}\tau^{m-1} + \cdots + v_1\tau + v_0$

Output: w =REGEN_TNAF(v)

$w \leftarrow v$;

$i \leftarrow 0$;

while ($w_j \neq 0$ for some $j \geq i$) **do**

  if ($w_i == 0$) **then**

    $i \leftarrow i + 1$;

  else

    $t_0 \leftarrow w_i$;

    $t_1 \leftarrow w_{i+1}$;

    $w_i \leftarrow 0$;

    $w_{i+1} \leftarrow 0$;

    $w \leftarrow w$+TNAF($t_1\tau + t_0$);

    $i \leftarrow i + 1$;

multiplications, inversions and squarings, ignoring the cost of addition. In the case of this architecture, field multiplication and squaring are completed quickly enough that the cost of field addition becomes significant. The run times using binary, binary NAF and τ-adic NAF representations are shown in Table 6. These values are based on the curve addition and doubling equations defined in (5) and (6) assuming arbitrary curve parameters α and β and the average Hamming weights discussed in the previous sections. For the case of τ-NAF, a Frobenius mapping is assumed to require three squaring operations. The symbols $\mathcal{M}$, $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{I}$ correspond to field multiplication, squaring, addition and inversion respectively. In each case it is assumed that the length of the integer is approximately equal to m.

5. A CO-PROCESSOR ARCHITECTURE FOR ECC SCALAR MULTIPLICATION

In the recent past, several articles have proposed various hardware architectures/ accelerators for ECC. These elliptic curve cryptographic accelerators can be categorized into three functional groups. They are

Table 6. Cost of Scalar Multiplication in terms of Field Operations

	Generic m	$m = 163$
Binary	$(10\mathcal{M} + 7\mathcal{S} + 8\mathcal{A})m + \mathcal{I}$	$1630\mathcal{M} + 1141\mathcal{S} + 1304\mathcal{A} + \mathcal{I}$
NAF	$(\frac{25}{3}\mathcal{M} + \frac{19}{3}\mathcal{S} + \frac{20}{3}\mathcal{A})m + \mathcal{I}$	$1359\mathcal{M} + 1033\mathcal{S} + 1087\mathcal{A} + \mathcal{I}$
τ-NAF	$(\frac{10}{3}\mathcal{M} + \frac{13}{3}\mathcal{S} + \frac{8}{3}\mathcal{A})m + \mathcal{I}$	$544\mathcal{M} + 706\mathcal{S} + 435\mathcal{A} + \mathcal{I}$

1. Accelerators which use general purpose processors to implement curve operations but implement the finite field operations using hardware. References [2] and [30] are examples of this. Both of these implementations support the composite field GF(2^{155}).

2. Accelerators which perform both the curve and field operations in hardware but use a small field size such as GF(2^{53}). Architectures of this type include those proposed in [28] and [8]. In [28], a processor for the field GF(2^{168}) is synthesized, but not implemented. Both works discuss methods to extend their implementation to a larger field size but do not actually do so.

3. Accelerators which perform both curve and field operations in hardware and use fields of cryptographic strength such as GF(2^{163}). Processors in this category include [3, 10, 17, 25, 27].

The work discussed in this section falls into category three. The architectures proposed in [25] and [27] were the first reported cryptographic strength elliptic curve co-processors. Montgomery scalar multiplication with an LSD multiplier was used in [27]. In [25] a new field multiplier is developed and demonstrated in an elliptic curve scalar multiplier. In both [17] and [3] parameterized module generation is discussed. To the best of our knowledge the architecture proposed in [10] offers the fastest scalar multiplication using FPGA technology at 0.144 milliseconds. This architecture uses Montgomery scalar multiplication with López and Dahab's projective coordinates. They use a shift and add field multiplier but also compare LSD and Karatsuba multipliers.

This section describes a hardware architecture for elliptic curve scalar multiplication. The architecture uses projective coordinates and is optimized for scalar multiplication over the Koblitz curves using the arithmetic routines discussed in Section 3 to perform the field arithmetic.

5.1. Co-processor Architecture

The architecture, which is detailed in this section, consists of several finite field arithmetic units, field element storage and control logic. All logic related to finite field arithmetic is optimized for specific field size and reduction polynomial. Internal curve computations are performed using López & Dahab's projective coordinate system.

While generic curves are supported, the architecture is optimized specifically for the special Koblitz curves.

The processor's architecture consists of the data path and two levels of control. The lower level of control is composed of a micro-sequencer which holds the routines required for curve arithmetic such as `DOUBLE` and `ADD`. The top level control is implemented using a state machine which parses the scalar and invokes the appropriate routines in the lower level control. This hierarchical control is shown in Figure 7.

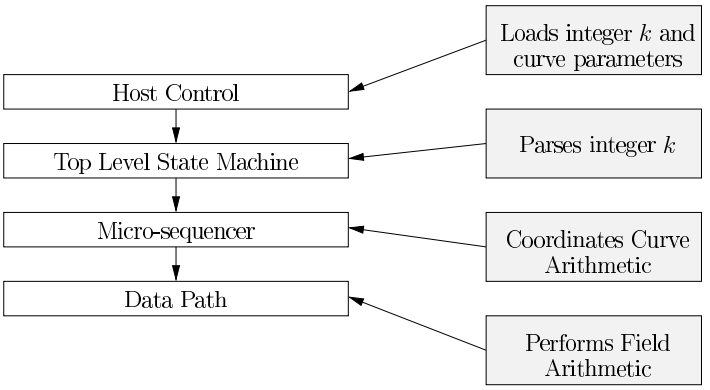

Figure 7. Co-Processor's Hierarchical Control Path

Co-processor Data Path

The data path of the co-processor consists of three finite field arithmetic units as well as space for operand storage. The arithmetic units include a multiplier, adder, and squaring unit. Each of these are optimized for a specific field and corresponding field polynomial. In an attempt to minimize time lost to data movement, the adder and multiplier are equipped with dual input ports which allow both operands to be loaded at the same time (the squaring unit requires a single operand and cannot benefit from an extra input bus). Similarly, the field element storage has two output ports used to supply data to the finite field units. In addition to providing field element storage, the storage unit provides the connection between the internal m-bit data path and the 32-bit external world. Figure 8 shows how the arithmetic units are connected to the storage unit.

The internal m-bit busses connecting the storage and arithmetic units are controlled to perform sequences of field operations. In this way the underlying curve operations `DOUBLE` and `ADD` as well as field inversion are performed.

Field Element Storage: The field element storage unit provides storage for curve points and parameters as well as temporary values. Parameters required to perform

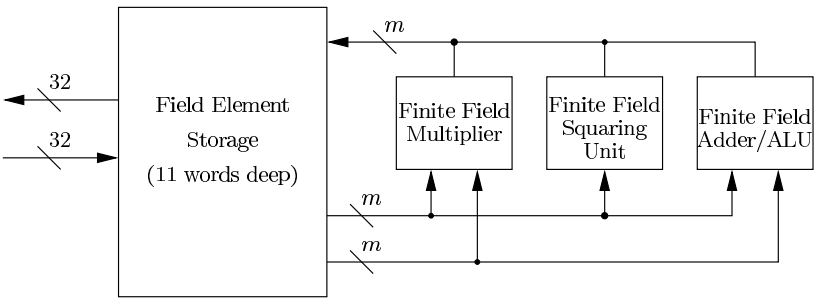

Figure 8. Co-Processor Data-Path

elliptic curve scalar multiplication include the field elements α and β and coordinates of the base point P. Storage will also be required for the coordinates of the scalar multiple Q. The point addition routine developed for this design also requires four temporary storage locations for intermediate values. Figure 9 shows how the storage space is organized.

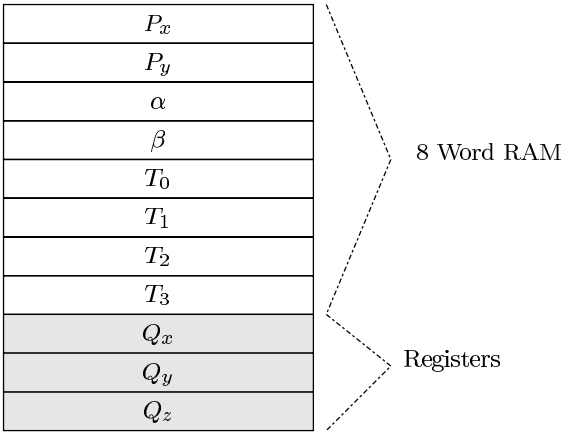

Figure 9. Field Element Storage

The top eight field element storage locations are implemented using 32-bit dual-port RAMs generated by the Xilinx Coregen tool and the bottom three storage locations[2]

[2] These locations are shaded gray in Figures 9 and 10.

are made of register files with 32-bit register widths. The dual 32-bit/m-bit interface support is achieved by instantiating $\lceil \frac{m}{32} \rceil$ dual-port storage blocks (either memories or register files) with 32-bit word widths as shown in Figure 10. The figure assumes $m = 163$. If the 32-bit storage locations in Figure 10 are viewed as a matrix then the rows of the matrix hold the m-bit field words. Each 32-bit location is accessible by the 32-bit interface and each m-bit location is accessible by the m-bit interface. For simplicity sake the field elements are aligned at 32 byte boundaries.

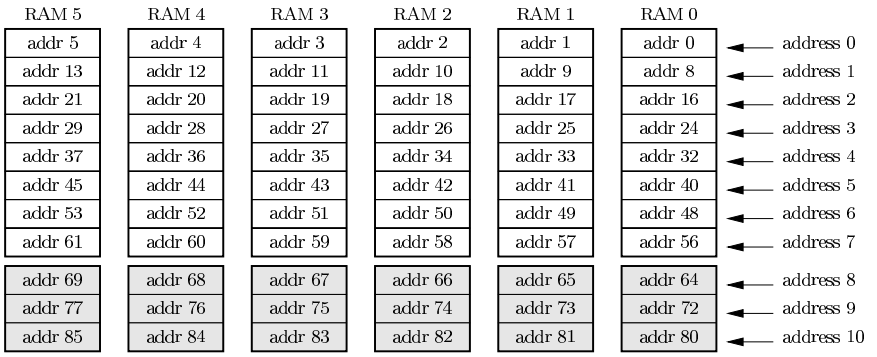

Figure 10. 32-bit/163-bit Address Map

Computation of τQ: In addition to providing storage, the registers in the bottom three m-bit locations are capable of squaring the resident field element. This is accomplished by connecting the logic required for squaring directly to the output of the storage register. The squared result is then muxed in to the input of the storage register and is activated with an enable signal. Figure 11 provides a diagram of this connection. This allows the squaring operations required to compute τQ to be performed in parallel. Furthermore, it eliminates the data movement otherwise required if the squaring unit were to be loaded and unloaded for each coordinate of Q. This provides significant performance improvement when using Koblitz curves.

The Micro-sequencer

The micro-sequencer controls the data movement between the field element storage and the finite field arithmetic units. In addition to the fundamental load and store operations, it supports control instructions such as jump and branch. The following list briefly summarizes the instruction set supported by the micro-sequencer.

- ld: Load operand(s) from storage location into specified field arithmetic unit.
- st: Store result from field arithmetic unit into specified storage location.
- j: Jump to specified address in the micro-sequencer.

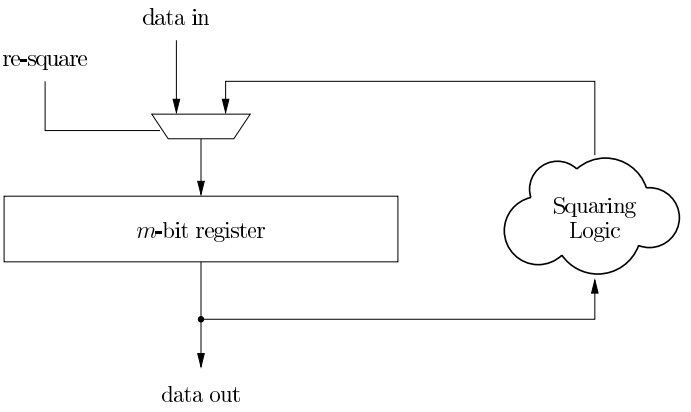

Figure 11. Efficient Frobenius Mapping

- jr: Jump to specified micro-sequencer address and push current address onto the program counter stack.
- ret: Return to micro-sequencer address. The address is supplied by the program counter stack.
- bne: Branch if the last field elements loaded into the ALU are NOT equal.
- nop: Increment program counter but do nothing.
- set: Set internal counter to specified value.
- rsq: Resquares the contents of the squaring unit.
- dbnz: Decrement internal counter and branch if the new value of the counter is zero. This opcode also causes the contents of the squaring unit to be resquared.

A two-pass perl assembler was developed to generate the micro-sequencer bit code. The assembler accepts multiple input files with linked addresses and merges them into one file. This file is then used to generate the bit code. The multiple input file support allows different versions of the ROM code to be efficiently managed. Different implementations of the same micro-sequencer routine can be stored in different files allowing them to be easily selected at compile time.

Micro-sequencer Routines: The micro-sequencer supports the curve arithmetic primitives, field inversion as well as a few other miscellaneous routines. The list below provides a summary of routines developed for use in the design.

- `POINT_ADD` (P, Q): Adds the elliptic curve points P and Q where P is represented in affine coordinates and Q is represented using projective coordinates. The result is given in projective coordinates.

- POINT_SUB (P, Q): Computes the difference $Q - P$. P is represented using affine coordinates and Q is represented using projective coordinates. The result is given in projective coordinates. This routine calls the POINT_ADD routine.
- POINT_DBL (Q): Doubles the elliptic curve point Q. Both Q and the result are in projective coordinates.
- INVERT (X): Computes the inverse of the finite field element X.
- CONVERT (Q): Computes the affine coordinates of an elliptic curve point Q given the point's projective coordinates. This routine calls the INVERT routine.
- COPY_P2Q (P, Q): Copies the x and y coordinates of point P to the x and y coordinates of point Q. The z coordinate of point Q is set to 1.
- COPY_MP2Q (P, Q): Computes the x and y coordinates of point $-P$ and copies them to the x and y coordinates of point Q. The z coordinate of point Q is set to 1.

Several versions of the POINT_ADD routine have been developed. The most generic one supports any curve over the field GF(2^m). In this version, the values of α and β are used when computing the sum of two points. This curve also checks if $Q \neq P$, $Q \neq -P$ and $Q \neq \mathcal{O}$. The second version of the point addition routine is optimized for a Koblitz curve by assuming α and β are equal to the NIST recommended values. The number of field multiplications required to compute the addition of two points is reduced from 10 to 9. The third version of the routine is optimized for a Koblitz curve and also forgoes the checks of point Q. If the base point P has a large prime order and the integer k is less than this order[3], it will never be the case that $Q = \pm P$ or $Q = \mathcal{O}$. This final version of the routine is the fastest of the three routines and is the one used to achieve the results reported at the end of the section.

Top Level Control

The routines listed above along with the POINT_FRB(Q) operation are invoked by the top level state machine. The POINT_FRB(Q) routine computes the Frobenius map of the point Q. This operation is not as complex as the other operations and is not implemented in the micro-sequencer. It is invoked by the top level state machine all the same.

The state machine parses the scalar k and calls the routines as needed. Since integers in NAF and τ-NAF require use of the symbol -1 (denoted $\bar{1}$), the scalar requires more than just an m-bit register for storage. In the implementation given here, each symbol in the scalar is represented using two bits; one for the magnitude and one for the sign. Table 7 provides the corresponding representation. For each bit k_i in the scalar k the magnitude is stored in the register $k_i^{(m)}$ and the sign is stored in register

[3] These are fair assumptions since the security of the ECC implementation relies on these properties.

Table 7. Representation of the Scalar k

Symbol	Magnitude	Sign
0	0	-
1	1	0
$\bar{1}$	1	1

$k_i^{(s)}$. Table 8 provides example representations for integers in binary form, NAF, and τ-adic NAF using $m = 8$.

Table 8. Example Representations of the Scalar

k	$k^{(m)}$	$k^{(s)}$
$(01001100)_2$	$(01001100)_2$	$(00000000)_2$
$(0100\bar{1}010)_{NAF}$	$(01001010)_2$	$(00001000)_2$
$(0100\bar{1}010)_{\tau-NAF}$	$(01001010)_2$	$(00001000)_2$

The top level state machine is designed to support binary, NAF and τ-adic NAF representations of the scalar. This effectively requires the state machine to perform Algorithms 3, 6 and 8. By taking advantage of the similarities between these algorithms, the top level state machine can perform this task with the addition of a single mode. This is shown in Algorithm 12. The algorithm is written in terms of the underlying curve and field primitives provided by the micro-sequencer (listed in Section 5.1).

The first step of Algorithm 12 is to search for the first non-zero bit in $k^{(m)}$. Once found, either P or $-P$ is copied to Q depending on the sign of the non-zero bit. The **while** loop then iterates over all the remaining bits in the scalar performing "doubles and adds" or "Frobenius mappings and adds" depending on the mode. Since the curve arithmetic is performed using projective coordinates, the result must be converted to affine coordinates at the end of computation.

Choice of Field Arithmetic Units

The use of redundant arithmetic units, specifically field multipliers, has been suggested in [3] and should be considered when designing an elliptic curve scalar multiplier.

Algorithm 12. State Machine Algorithm

Inputs: $k^{(m)} = (k_{l-1}^{(m)}, k_{l-2}^{(m)}, \ldots, k_1^{(m)}, k_0^{(m)})_2$,
$k^{(s)} = (k_{l-1}^{(s)}, k_{l-2}^{(s)}, \ldots, k_1^{(s)}, k_0^{(s)})_2$,
Point P and *mode* (NAF or τ-NAF)

Output: Point $Q = kP$

$i \leftarrow l - 1$;
while $(k_i^{(m)} == 0)$ **do**
  $k \leftarrow i - 1$;
if $(k_i^{(s)} == 1)$ **then**
  COPY_MP2Q(P, Q);
else
  COPY_P2Q(P, Q);
$i \leftarrow i - 1$;
while $(i \geq 0)$ **do**
  if (*mode* == τ-NAF) **then**
    $Q \leftarrow$ POINT_FRB(Q);
  else
    $Q \leftarrow$ POINT_DBL(Q);
  if $(k_i^{(m)} == 1)$ **then**
    if $(k_i^{(s)} == 1)$ **then**
      $Q \leftarrow$ POINT_SUB(Q, P);
    else
      $Q \leftarrow$ POINT_ADD(Q, P);
  $i \leftarrow i - 1$
$Q \leftarrow$ CONVERT(Q);

It seems the advantage provided remains purely theoretical. This can be seen by examining the top performing ECC multipliers in [10] and [27], both of which use a single field multiplier. Reasons for doing the same for this ECC accelerator are twofold. (1) One of the limiting factors for the performance of the design is data movement. As shown in Figures 12 and 13 the bus usage for point addition and point doubling is very high (83% and 80% respectively). If another multiplier is added to the design there may not be enough free bus cycles to capitalize on the extra computational power. For the field GF(2^{163}), the multiplier computes a product in four clock cycles and requires three cycles to load and unload the unit. If a second multiplier is added, then two multiplications can be completed in four cycles but six cycles are required to unload the multiplier. (2) Many of the multiplications in point addition and point doubling are dependent on each other and must be performed in sequence. For this reason, the second multiplier may sit idle much of the time. The combination of these observations seems to argue against the use of multiple field multiplication units in the design.

5.2. FPGA Prototype

A prototype of the architecture has been implemented for the field GF(2^{163}) using the NIST recommended field polynomial. The design was coded using Verilog HDL and synthesized using Synopsys FPGA Compiler II. Xilinx' Foundation software was used to place, route and time the netlist. The prototype was designed to run at 66 MHz on a Xilinx' Virtex 2000E FPGA.

The resulting design was verified on the Rapid Prototyping Platform (RPP) provided by Canadian Microelectronics Corporation (CMC) [5, 6]. The hardware/software system includes an ARM Integrator/LM-XCV600E+ (board with a Virtex 2000E FPGA) and an ARM Integrator/ARM7TDMI (board with an ARM7 core) connected by the ARM Integrator/AP board. The design was connected to an AHB slave interface which made it directly accessible by the ARM7 core. Stimulated by compiled C-code, the core read from and wrote to the prototype. The Integrator/AP's system clock had a maximum frequency of 50 MHz. In order to run our design at 66 MHz it was necessary to use the oscillator generated clock provided with the Integrator/LM-SCV600E+. The data headed to and coming from the design was passed across the two clock domains.

5.3. Results

Table 9 shows the performance in clock cycles of the prototypes field and curve operations. These values were gathered using a field multiplier digit size of $g = 41$.

Note that the multiple instantiations of the squaring logic allow for the Frobenius mapping of a projective point to be completed in a single cycle. This significantly improves the performance of scalar multiplication when using the Koblitz curves.

The prototype of the scalar multiplier has been implemented using several digit sizes in the field multiplier. Table 10 reports the area consumption and resulting performance of the architecture given the different digit sizes. Table 11 provides a comparison of published performance results for scalar multiplication.

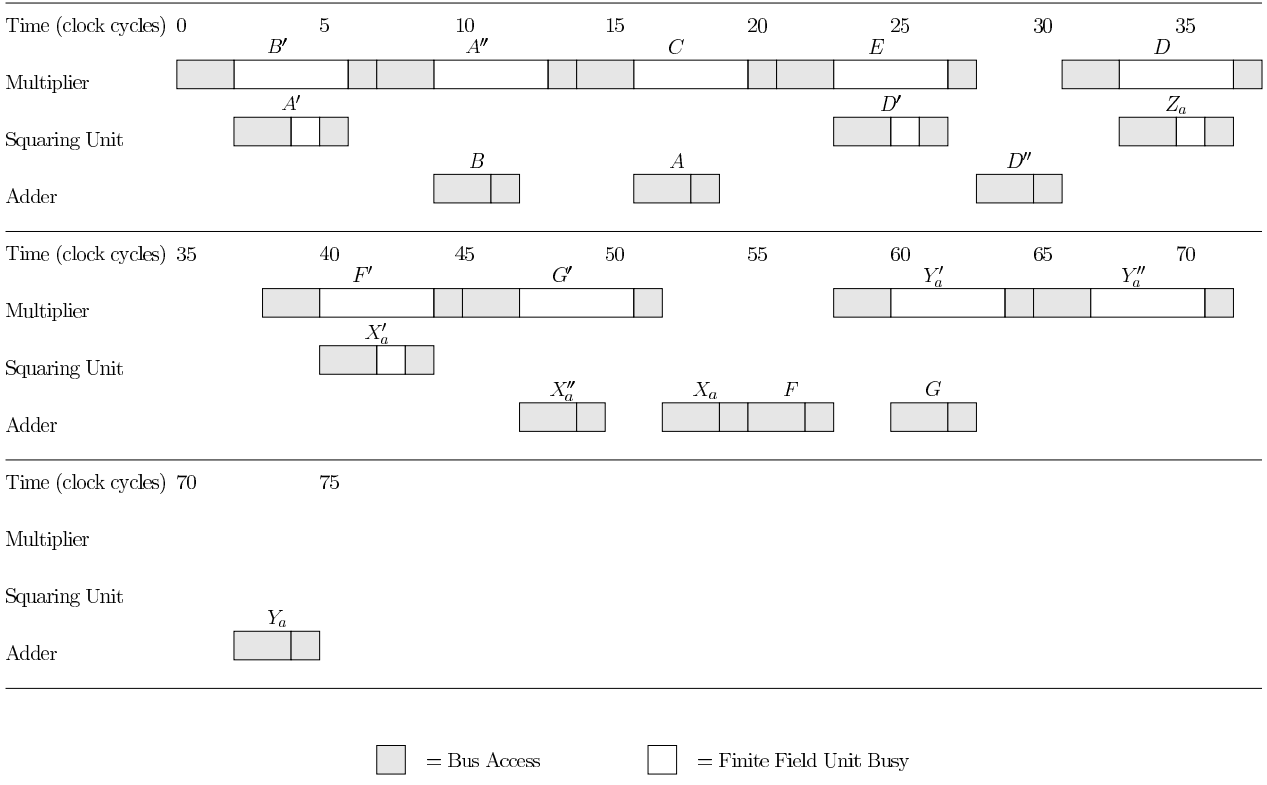

Figure 12. Utilization of Finite Field Units for Point Addition

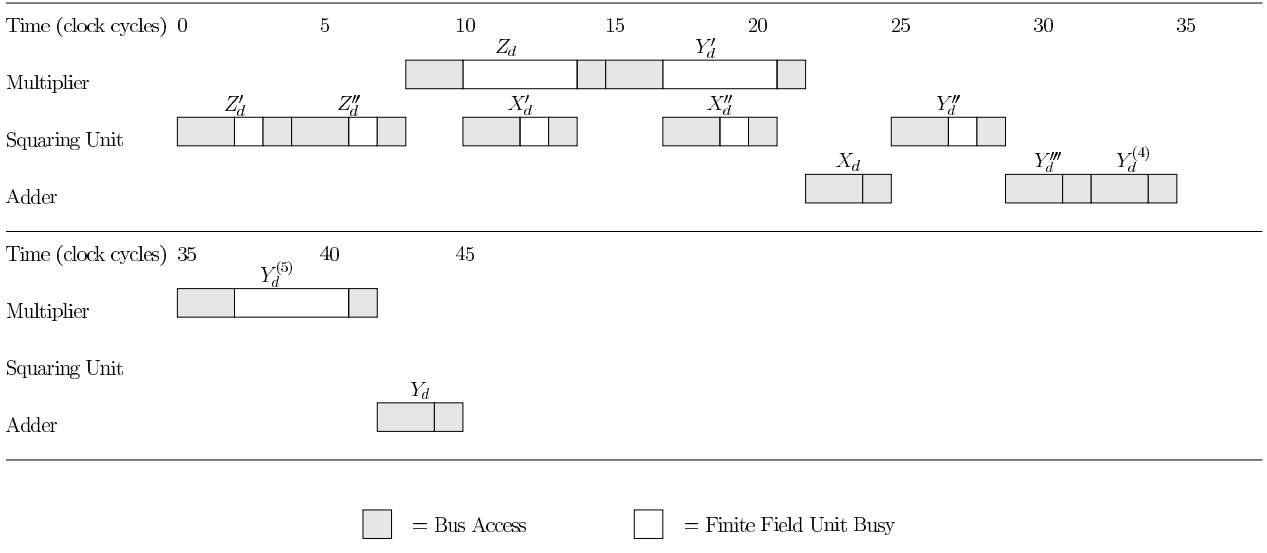

Figure 13. Utilization of Finite Field Units for Point Doubling

Table 9. Performance of Field and Curve Operations

Operation ($g = 41$)	# Cycles
Point Addition	79
Point Subtraction	87
Point Double	68
Frobenius Mapping	1

Table 10. Performance and Cost Results for Scalar Multiplication

Multiplier Digit Size	# LUTs	# FFs	Binary (ms)	NAF (ms)	τ-NAF (ms)
$g = 4$	6,144	1,930	1.107	0.939	0.351
$g = 14$	7,362	1,930	0.446	0.386	0.135
$g = 19$	7,872	1,930	0.378	0.329	0.113
$g = 28$	8,838	1,930	0.309	0.272	0.090
$g = 33$	9,329	1,930	0.286	0.252	0.083
$g = 41$	10,017	1,930	0.264	0.233	0.075

Table 11. Comparison of Published Results

Implementation	Field	FPGA	Scalar Mult. (ms)
S. Okada et. al. [25]	GF(2^{163})	Altera EPF10K250	45
Leong & Leung [17]	GF(2^{155})	Xilinx XCV1000	8.3
M. Bednara et. al. [3]	GF(2^{191})	Xilinx XCV1000	0.27
Orlando & Paar [27]	GF(2^{167})	Xilinx XCV400E	0.210
N. Gura et. al. [10]	GF(2^{163})	Xilinx XCV2000E	0.144
Our design ($g = 14$)	GF(2^{163})	Xilinx XCV2000E	0.135
Our design ($g = 41$)	GF(2^{163})	Xilinx XCV2000E	0.075

The performance of 0.144 ms reported in [10] is the fastest reported scalar multiplication using FPGA technology. The design presented in this report provides almost double (0.075 ms) the performance for the specific case of Koblitz curves.

The co-processor discussed in this chapter requires approximately half the CLBs used in the co-processor of [10] using the same FPGA. It must be noted that the co-processor presented in [10] is robust in that it supports all fields up to GF(2^{256}). In applications where support for a only single field size is required it is overkill to support elliptic curves over many fields. In scenarios such as this, this new elliptic curve co-processor offers an improved cost effective solution.

6. CONCLUDING REMARKS

In this chapter, the development of an elliptic curve cryptographic co-processor has been discussed. The co-processor takes advantage of multiplication and squaring arithmetic units which are based on the look-up table-based multiplication algorithm proposed in [11]. Field elements are represented with respect to the polynomial basis. While the base point and resulting scalar are given in affine coordinates, internal arithmetic is performed using projective coordinates. This choice of coordinate system allows the scalar multiple of a point to be computed with a single field inversion alleviating the need for a highly efficient inversion method. The processor was designed to support signed, unsigned and τ-NAF integer representation. All curves over a specific field are supported, but the architecture is optimized specifically for the Koblitz curves.

7. ACKNOWLEGEMENTS

This work was supported in part by the Security Technology Center in the Semiconductor Products Sector of Motorola. Dr. Hasan's work was supported in part by NSERC. Pieces of the work were presented at SPIE 2003 [19] and ITCC 2004 [20].

8. REFERENCES

1. *Wireless Application Protocol - Version 1.0*, 1998.
2. G. B. Agnew, R.C. Mullin, and S. A. Vanstone. An implementation of elliptic curve cryptosystems over $F_{2^{155}}$. *IEEE Journal on Slected Areas in Communications*, 11:804–813, June 1993.
3. Marcus Bednara, Michael Daldrup, Joachim von zur Gathen, Jamshid Shokrollahi, and Jurgen Teich. Implementation of elliptic curve cryptographic coprocessor over GF(2^m) on an FPGA. In *International Parallel and Distributed Processing Symposium: IPDPS Workshops*, April 2002.
4. D. Chudnovsky and G. Chudnovsky. Sequences of numbers generated by addition in formal groups and new primality and factoring tests. *Advances in Applied Mathematics*, 1987.
5. Canadian Microelectronics Corporation. *CMC Rapic-Prototyping Platform: Design Flow Guide*, 2002.
6. Canadian Microelectronics Corporation. *CMC Rapic-Prototyping Platform: Installation Guide*, 2002.
7. T. Dierks and C. Allen. *The TLS Protocol - Version 1.0 IETF RFC 2246*, 1999.

8. Lijun Gao, Sarvesh Shrivastava, and Gerald E. Sobelman. Elliptic curve scalar multiplier design using FPGAs. In *Cryptographic Hardware and Embedded Systems (CHES)*, 1999.

9. Daniel M. Gordon. A survey of fast exponentiation methods. *J. Algorithms*, 27(1):129–146, 1998.

10. Nils Gura, Sheueling Chang Shantz, Hans Eberle, Summit Gupta, Vipul Gupta, Daniel Finchelstein, Edouard Goupy, and Douglas Stebila. An end-to-end systems approach to elliptic curve cryptography. In *Cryptographic Hardware and Embedded Systems (CHES)*, 2002.

11. M. Anwarul Hasan. Look-up table-based large finite field multiplication in memory constrained cryptosystems. *IEEE Transactions on Computers*, 49(7), July 2000.

12. IEEE. *P1363: Editorial Contribution to Standard for Public Key Cryptography*, February 1998.

13. T. Itoh and S. Tsujii. A fast algorithm for computing multiplicative inverses in GF(2^m) using normal bases. *Information and Computing*, 78(3):171–177, 1988.

14. Brian King. An improved implementation of elliptic curves over GF(2^n) when using projective point arithmetic. In *Selected Areas in Cryptography*, 2001.

15. Neal Koblitz. Elliptic curve cryptosystems. *Mathematics of Computation*, 1987.

16. Neal Koblitz. CM curves with good cryptographic properties. In *Advances in Cryptography, Crypto '91*, pages 279–287. Springer-Verilag, 1991.

17. Philip H. W. Leong and Ivan K. H. Leung. A microcoded elliptic curve processor using FPGA technology. *IEEE Transactions on VLSI Systems*, 10(5), October 2002.

18. Julio Lopez and Ricardo Dahab. Improved algorithms for elliptic curve arithmetic in GF(2^n). In *Selected Areas in Cryptography*, pages 201–212, 1998.

19. Jonathan Lutz and Anwarul Hasan. High performance finite field multiplier for cryptographic applications. In *SPIE's Advanced Signal Processing Algorithms, Architectures, and Implemenations*, Volume 5205, pages 541-551, 2003.

20. Jonathan Lutz and Anwarul Hasan. High performance fpga based elliptic curve cryptographic coprocessor. In *IEEE International Conference on Information Technology (ITCC)*, Volume II, pages 486-492, 2004.

21. Alfred Menezes. Elliptic curve public key cryptosystems. *Kluwer Academic Publishers*, 1993.

22. A. Menezes, E. Teske, A. Weng. Weak Fields for ECC. Technical Report CORR 2003-15, Centre for Applied Cryptographic Research, University of Waterloo, 2003. *See http://www.cacr.math.uwaterloo.ca*

23. Victor Miller. Uses of elliptic curves in cryptography. In *Advances in Cryptography, Crypto '85*, 1985.

24. NIST. *FIPS 186-2 draft, Digital Signature Standard (DSS)*, 2000.

25. Souichi Okada, Naoya Torii, Kouichi Itoh, and Masahiko Takenaka. Implementation of elliptic curve cryptographic coprocessor over GF(2^m) on an FPGA. In *Cryptographic Hardware and Embedded Systems (CHES)*, pages 25–40. Springer-Verlag, 2000.

26. OpenSSL. *See http://www.openssl.org*.

27. Gerardo Orlando and Christof Paar. A high-performance reconfigurable elliptic curve processor for GF (2^m). In *Cryptographic Hardware and Embedded Systems (CHES)*, 2000.

28. Martin Christopher Rosner. Elliptic curve cryptosystems on reconfigurable hardware. Master's thesis, Worcester Polytechnic Institute, 1998.

29. Jerome A. Solinas. Improved algorithms for arithmetic on anomalous binary curves. In *Advances in Cryptography, Crypto '97*, 1997.

30. S. Sutikno, R. Effendi, and A. Surya. Design and implemntation of arithmetic processor $F_{2^{155}}$ for elliptic curve cryptosystems. In *IEEE Asia-Pacific Conference on Circuits adn Systems*, pages 647–650, November 1998.