

SIMD Acceleration of Modular Arithmetic on Contemporary Embedded Platforms

Krishna Chaitanya Pabbuleti, Deepak Hanamant Mane, Avinash Desai, Curt Albert and Patrick Schaumont

Department of Electrical and Computer Engineering

Virginia Polytechnic and State University

Blacksburg, VA 24060

Email: {kriscp4, mdeepak, aviraj, falbert9, schaum}@vt.edu

Abstract—Elliptic curve cryptography (ECC) is a public key crypto system popular for embedded implementations because of its shorter key sizes. ECC computations are complex; they involve point additions and doublings on elliptic curves over finite fields. The execution time of ECC is completely dominated by modular multiplications in these fields. In this contribution, we propose vector processing techniques to accelerate modular multiplications in prime fields. We demonstrate implementations for the Venom (NEON) coprocessor in Qualcomm’s Scorpion (ARM) CPU, as well as for the SSE2 instruction-set extensions in Intel’s Atom CPU. Our implementations, which use NIST-standard prime-field curves, run more than two times faster than the OpenSSL versions of the same ECC operations on the same processor.

I. INTRODUCTION

Handheld computing is an emerging market with increasingly complex and powerful processors to provide wide range of services to the end user. Current processors targeted at the handheld market contain several coprocessors and accelerators which can perform specialized functions, such as signal processing and video acceleration, without burdening the main processor. Information security is a common requirement for those platforms as well. Public key cryptographic algorithms like RSA, DSA and elliptic curve cryptography (ECC) are needed to support key exchange and signature protocols. ECC in particular is popular in embedded context because of the smaller key sizes compared to other public key cryptographic systems. For example, at 128-bit equivalent security, we need 3072-bit public key in RSA, whereas ECC requires only a 256-bit public key.

Our work focuses on efficient implementation of modular arithmetic which forms the basis for ECC, by exploiting the capabilities of SIMD coprocessors in Intel Atom and Qualcomm Snapdragon processors. High-performance implementations are important, also for handheld computing platforms. Indeed, faster cryptography enables higher levels of equivalent security, or higher availability of handheld computing resources for other tasks. Furthermore, new applications such as e-cash and privacy-friendly attributes make extensive use of public-key primitives.

ECC was independently proposed by Neal Koblitz and Victor S. Miller and is based on point algebra of elliptic curves over a finite field. The security of ECC is based on the Elliptic Curve Discrete Logarithm Problem which states that given two points P and Q on the curve such that $Q = k.P$, it is very hard to find k . This operation is the basis for secure

Algorithm 1 Right-to-left binary method for point multiplication [8]

INPUT: $k = (k_{t-1}, \dots, k_1, k_0)_2$, $P \in E(F_q)$.

OUTPUT: $k.P$.

1. $Q \leftarrow \infty$.
 2. For i from 0 to $t-1$ **do**
 - 2.1 If $k_i = 1$ then $Q \leftarrow Q + P$.
 - 2.2 $P \leftarrow 2P$.
 3. Return(Q).
-

protocols for signing and key exchange: the private key k is a very long integer chosen at random and the public key is derived by scalar multiplication of point P with k . The scalar multiplication thus forms the basis of ECC. All points on an elliptic curve form a group. One can either add two different points, or one can double a point (add it to itself). Using point adding and point doubling, a very basic implementation of $k.P$ is shown in Algorithm 1. Each point operation (adding or doubling) requires multiple modular-arithmetic operations in the underlying field. For example, for a Weierstrass curve with Jacobian coordinates (a common choice for NIST curves), one may find 7 modular multiplications and 3 squaring operations for each point doubling, and 12 modular multiplications and 2 modular squarings for point adding [8]. It’s easy to see that computing $k.P$ is dominated by modular multiplications. If we assume a 192-bit random k , we may expect to see about 190 point-double and 85 point-add operations, leading to 1362 modular multiplications and 740 modular squarings for each $k.P$.

We propose efficient implementation techniques for modular multiplication using advanced architectural features (and SIMD in particular) on modern SoCs. Yan and others showed that DSP processors are efficient in accelerating modular arithmetic [15], [16]. Morozov later demonstrated how embedded DSP cores in SoC could be used to accelerate modular arithmetic [13]. Similarly, the SSE extension on x86 processors enables efficient big number multiplication [11]. Bernstein et al showed that significant performance gain can be achieved by using NEON SIMD extension [3]. However, his demonstration used a specialized curve called `curve25519`. In this work, we target NIST recommended curves over prime fields. NIST primes are special primes which are of the form $2^m \pm 2^n - \dots - 1$. Reduction is very easy in these fields because of the structure of the prime number. The five 5 NIST primes (P192, P224, P256, P384, P512) are used

TABLE I
COMPARISON BETWEEN INTEL AND SNAPDRAGON HARDWARE
PLATFORMS

Specification	Intel Atom N2800	Qualcomm Snapdragon APQ8060
Processor Speed	1.86 GHz	1 GHz
Number of Cores	2	2
Instruction Set	x86-64	ARMv7
Instruction Set Extensions	SSE2	Thumb2, Neon, VFPv3
Number of Registers	16 32-bit Registers in 32-bit mode, 16 64-bit Registers in 64-bit mode	16 32-bit registers (r0-r15)
Number of vector registers	16 128-bit registers (XMM0-XMM15)	16 128-bit registers (q0-q15)
L1 Cache Size	56 KB (24 KB data and 32 KB Instruction) with 2 cycle latency	16 KB data cache and 16 KB instruction cache
L2 Cache Size	1 MB (512 KB per core) with 15 cycle latency	512 KB

for standard curves called $nistp_{192}$, $nistp_{224}$, ..., which require modular operations using 192-bit, 224-bit, ... prime field numbers. Recent work by Kasper demonstrated portable (non-SIMD), 64-bit optimized versions of $nistp_{224}$ [17]; our work explores the opportunities offered by SIMD processor extensions.

We implement high performance modular arithmetic on two different hardware platforms supporting SIMD instructions and we compare their performance. We evaluate our implementation on Qualcomm Scorpion and on Intel Atom processor. The Scorpion processor is one of the seven processors integrated in Qualcomm’s Snapdragon platform; we will refer to the more common term Snapdragon for the remainder of this paper. We evaluate the performance of $nistp_{192}$ and $nistp_{224}$ elliptic curves on both the platforms. The same concept could be extended to other standard curves as well.

The major contributions of this work are:

- Implementation of efficient modular multiplication on SIMD architecture for NIST primes.
- Apples-to-apples performance comparison of vectorized modular arithmetic on contemporary embedded platforms, including cycle count performance and analysis of the instruction set.

The remainder of the paper is organized as follows. In the next section we describe the SIMD features of the Qualcomm Snapdragon and the Intel Atom. Section 3 explains different implementation techniques for modular multiplication and optimization for both hardware platforms. Section 4 shows implementation results, including measured performance for P192 and P224 and Section 5 concludes the paper.

II. HARDWARE PLATFORM

We compare the performance of modular arithmetic on two different hardware platforms namely, Qualcomm Snapdragon APQ8060 and Intel Atom N2800. The Snapdragon features a dual CPU (Scorpion) architecture, running up to 1.7 GHz each,

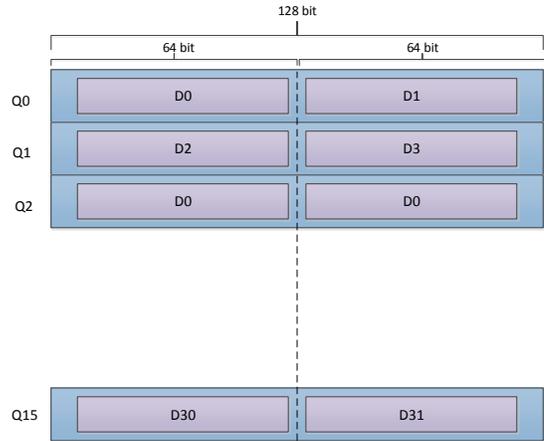


Fig. 1. Registers in NEON

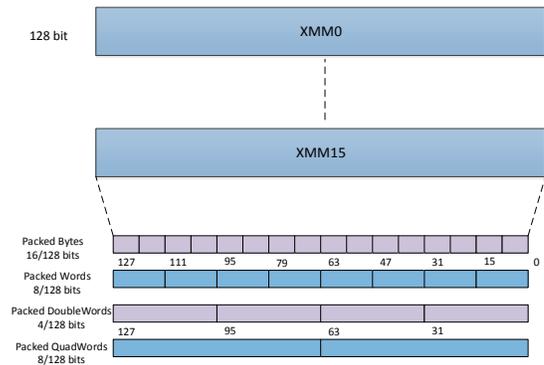
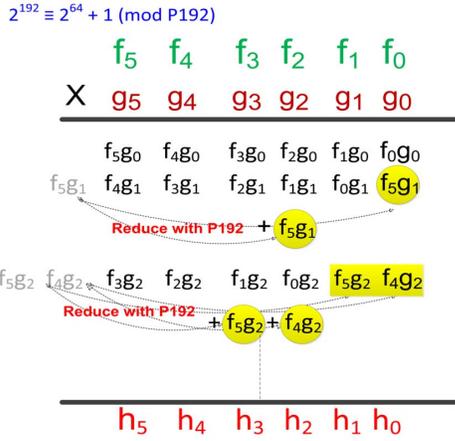


Fig. 2. Fundamental 128-Bit packed SIMD data types in SSE2 [6]

two VeNum (NEON) 128-bit SIMD multimedia coprocessors and combined 512 KByte L2 cache. The Scorpion processor is designed in-house but has many architectural similarities with the ARM Cortex-A8 and ARM Cortex-A9 CPUs. Functionally, Scorpion is an intermediary step between Cortex-A8 and Cortex-A9, supporting some but not all of Cortex-A9’s out-of-order instruction execution capabilities. Scorpion-based Snapdragon SOCs implement Cortex-A8 and A9-compliant floating-point and NEON SIMD engines. We specifically target NEON coprocessor for our SIMD optimizations.

The NEON architecture has sixteen 128-bit vector registers, q_0 through q_{15} as shown in Figure 1. It also includes thirty-two 64-bit vector registers, d_0 through d_{31} , but these registers share physical space with the 128-bit vector registers: q_0 is the concatenation of d_0 and d_1 , q_1 is the concatenation of d_2 and d_3 , and so on. The basic ARM architecture has only sixteen 32-bit registers, r_0 through r_{15} . Register r_{13} is the stack pointer and register r_{15} is the program counter, leaving only fourteen 32-bit registers for general use. An obvious benefit of NEON for cryptography is that it provides much more space in registers, reducing the number of loads and stores from memory. The 128-bit arithmetic unit can perform four 32-bit operations in each cycle. The Cortex A8 NEON microarchitecture has one 128-bit arithmetic unit and one 128-bit NEON load/store unit that runs in parallel with the NEON arithmetic unit.

The second processor we use is the Intel Atom N2800 Pro-



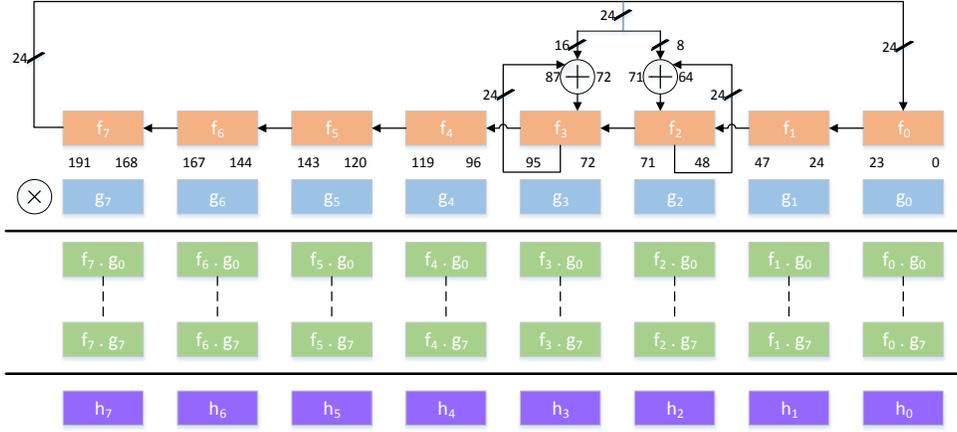


Fig. 4. P192 Implementation

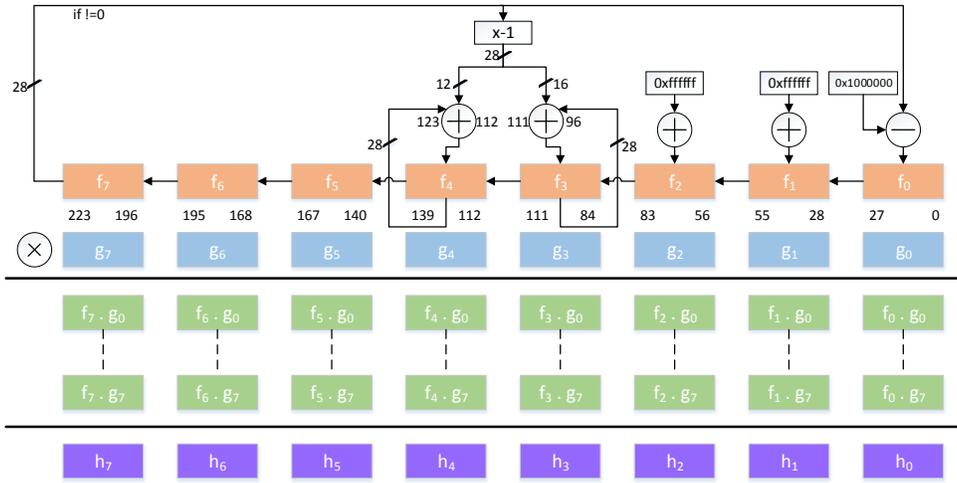


Fig. 5. P224 Implementation

2) *Multiplicand reduction method*: Instead of the shifting and reducing the partial products, one of the multiplicands is reduced after calculating one row of partial product. Mane discusses an FPGA implementation of modular multiplication in which one of the operands is reduced as the multiplication progresses [12]. We have adapted this technique on SIMD platforms by reducing the number of reduction stages. We have also eliminated intermediate carry propagation by using redundant representation of operands. The operands f and g are represented in redundant representation in radix-24 in the same way as discussed in the previous method. First row of partial products are obtained by multiplying f with g_0 . Now, f is shifted left by 24 bits, the 24 bits beyond 2^{192} are added to 2^0 and 2^{64} , which produces reduced multiplicand. This shifting and adding back is based on the relation $2^{192} = 2^{64} + 1$ in P192 prime field, which produces the reduced result. It should be noted that, because of the redundant representation, lower 8 bits of 2^{64} are in f_2 and remaining 16 bits are present in f_3 . So, lower 8 bits of f_7 are added in upper 8 bits of f_2 and upper 16 bits of f_7 are added in lower 16 bits of f_3 . We do not need to worry about carry because of the slack present due to redundant representation. Now, the second row of partial products is calculated by multiplying reduced f with g_1 and added to the previous partial products. Similarly, all partial product rows are computed by shifting f and reducing

it, followed by a multiplication with g . In the end, carries are propagated and any trailing bits beyond 2^{192} are again reduced by adding them to 2^0 and 2^{64} .

B. Modular Multiplication in P224

The next bigger prime in NIST primes is $P224 = 2^{224} - 2^{96} + 1$. Any number larger than this prime can be reduced by using the relation $2^{224} = 2^{96} - 1 \pmod{P224}$. Results of P192 indicate that the multiplicand reduction method is significantly faster than the schoolbook multiplication with intermediate reduction method. Therefore, we have implemented only multiplicand reduction method for P224.

1) *Multiplicand reduction method*: In the same way as above, one of the multiplicands is shifted and reduced after calculating one row of partial product. The numbers f and g are represented in redundant representation in radix-28 as in Listing 4

$$\begin{aligned}
 f &= f_0 \cdot 2^0 + f_1 \cdot 2^{28} + f_2 \cdot 2^{56} + f_3 \cdot 2^{84} + f_4 \cdot 2^{112} + f_5 \cdot 2^{140} + f_6 \cdot 2^{168} \\
 &\quad + f_7 \cdot 2^{196} \\
 g &= g_0 \cdot 2^0 + g_1 \cdot 2^{28} + g_2 \cdot 2^{56} + g_3 \cdot 2^{84} + g_4 \cdot 2^{112} + g_5 \cdot 2^{140} + g_6 \cdot 2^{168} \\
 &\quad + g_7 \cdot 2^{196}
 \end{aligned}$$

Listing 4. Redundant representation in Radix-28

First row of partial products are obtained by multiplying f with g_0 . Now, f is shifted left by 28 bits, the 28 bits beyond 2^{224} are added to 2^{96} and subtracted from 2^0 , which produces the reduced result. In the same way as above, because of the redundant representation, the lower 16 bits of 2^{96} are in f_3 and the remaining 12 bits are present in f_4 . The lower 16 bits of f_7 are added in upper 16 bits of f_3 and the upper 12 bits of f_7 are added in lower 12 bits of f_4 . When f is shifted left, f_0 is zero and we need to subtract f_7 from f_0 . We need to get borrow from neighbor, but the neighbor is not guaranteed to be a non-zero number. The borrow is taken from f_7 which is being added to f_3 and f_4 . Because we have taken the borrow from non-neighbor, we have to add $0xffffffff$ to the members between f_3 and f_0 , i.e., f_1 and f_2 . This is best illustrated in the Figure 5.

C. Platform Specific Optimization

This section explains specific implementation details for NEON and SSE2 vector coprocessors. NEON can do two multiplications or two multiply and accumulate operations at a time. Moreover, we can take advantage of multiply and accumulate instruction. As shown in Figures 4 and 5, we first multiply f_0 and f_1 with g_0 in one cycle and obtain partial products f_0g_0 and f_1g_0 . In the next cycle, we multiply f_2 and f_3 with g_0 and obtain f_2g_0 and f_3g_0 and so on. After multiplying all the limbs of f with g_0 , f is shifted left by 24-bits and reduced. It should be noted that this reduction is performed on main processor because it is more efficient in handling these operations compared to Neon coprocessor. In the next step, f_0 and f_1 are multiplied with g_1 and the results are added to previously calculated f_0g_0 and f_1g_0 , using multiply accumulate instruction. This gives additions for free and also saves registers, as there is no need to save the intermediate partial products. In the end, the partial products are moved to main processor, carry propagation is done and final product is calculated.

In SSE2, the modular multiplication is implemented as follows. We use `PMULUDQ` to perform two multiplication operations simultaneously. `PMULUDQ` can perform multiplication operation on two of the four 32-bit values read from memory. We can use `PSHUFD` to rearrange the 32-bit values to allow us to use all of the data read in from memory before writing it back or reading in new data. `PADDQ` is used to accumulate the 64-bit partial products produced in the multiplicand reduction algorithm. `PSHUFD` instruction is used to order the result so that it can be properly written back to memory, 128 bits at a time. The downside of the SSE2 instructions is that they overwrite one of the operands. When you call an instruction on two operands, it calculates the result and stores it in one of the operand registers overwriting the operand. In order to preserve the operand, we have store it in a different register beforehand which adds overhead. Moreover, SSE2 doesn't have multiply and accumulate instruction, so the intermediate results have to be saved and added to the partial product, adding further overhead.

IV. RESULTS

Table II shows the performance for the two proposed approaches on the Qualcomm Snapdragon and Intel Atom for prime field P192. The code is compiled using gcc version

TABLE II
CYCLES FOR MODULAR MULTIPLICATION FOR P192

Algorithm	Snapdragon	Snapdragon +NEON	Atom	Atom +SSE2
Intermediate Reduction Method	1116	858	1282	842
Multiplicand Reduction Method	574	404	864	685
GMP	1786	-	2122	-

TABLE III
CYCLES FOR MODULAR MULTIPLICATION FOR P224

Algorithm	Snapdragon	Snapdragon +NEON	Atom	Atom +SSE2
Multiplicand Reduction Method	685	405	1083	836
GMP	1858	-	2286	-

4.4.5 on ARM and gcc version 4.6.3 on Intel Atom. As our modular multiplication is coded in assembly, we expect that the difference in compiler versions has minimal impact. Clock cycles are measured by reading counter registers from Performance Monitoring Unit inside CP15 coprocessor of ARM. On Intel, clock cycles are measured using `RDTSC` instruction. Schoolbook multiplication is significantly slower than the multiplicand reduction method, because the former method involves several shifts and adds which adds a lot of overhead. Atom is slower than ARM in all the cases. Similar trend can be noticed in the numbers obtained from eBACS benchmark in Table IV [4]. Vectorising the multiplications increases the performance by almost 30% in all the cases. Both our methods are faster than the modular multiplication using GMP multiprecision library. Table III shows clock cycles comparison for multiplicand reduction method on the ARM and Atom for prime field P224. We did not implement schoolbook multiplication for P224 because results from P192 show that multiplicand reduction is already faster. Vectorising the modular multiplication gives a 40% improvement in the performance on ARM whereas gives only 22% improvement on Intel Atom. This may be attributed to overheads present in SSE2 like non-availability of multiply-accumulate instruction and result replacing one of the operands.

Table IV shows the comparison of point multiplication on various embedded platforms against point multiplication using our modular multiplication techniques. The point multiplication is performed in projective coordinate system in all the implementations given in the table. Our implementation is faster than the other generic implementations of comparable security level. However, specialized implementations like `curve25519` on Snapdragon by Bernstein [3] performs better than our implementation. As expected, Atom takes a larger number of cycles than Snapdragon because the modular multiplication on Atom takes more number of cycles. A similar trend is found in the cycle counts obtained from the eBACS benchmark suite, where Atom is 1.5 to 1.7 times slower than Snapdragon.

Figure 6 shows the data from Table IV in a manner to appreciate the merit of SIMD vectorization. The figure shows the number of Point Multiplications completed per million cycles of the target architecture, for different curve sizes and different architectures. The vectorized implementations (SSE2

TABLE IV
SCALAR MULTIPLICATION PERFORMANCE FOR VARIOUS EMBEDDED PLATFORMS

Platform	Curve and Implementation	Cycles
TI OMAP 3530(ARM A8 + C64x + DSP) [13]	secp160r1	1,054K
TI OMAP 3530(ARM A8 + C64x + DSP) [13]	secp224r1	2,175K
Qualcomm Snapdragon	Ed25519 [4]	3,295K
Intel N280	Ed25519 [4]	5,774K
Qualcomm Snapdragon	P192 ebacs_OpenSSL	3,143K
Qualcomm Snapdragon	P224 ebacs_OpenSSL	3,996K
Intel N280	P192 ebacs_OpenSSL	4,973K
Intel N280	P224 ebacs_OpenSSL	6,349K
Qualcomm Snapdragon with NEON	Curve25519 [3]	511K
Qualcomm Snapdragon	P192 with Multiplicand Reduction	1,591K
Qualcomm Snapdragon with NEON	P192 with Multiplicand Reduction	1,243K
Intel N280	P192 with Multiplicand Reduction	2,390K
Intel N280 with SSE2	P192 with Multiplicand Reduction	2,115K

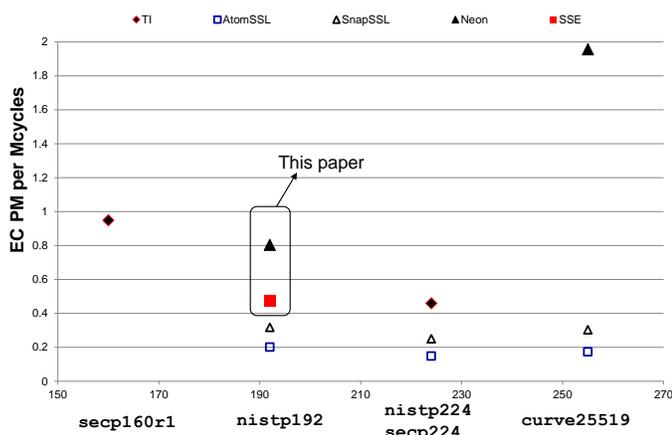


Fig. 6. Number of Point Mult per million Cycles

on Atom, NEON on Snapdragon, DSP64x on OMAP) are shown as solid symbols; the default ones are shown as open symbols.

V. CONCLUSION

In this paper we demonstrated how SIMD extensions could be exploited to accelerate the underlying modular arithmetic. We implemented and compared the modular multiplication on two different embedded computing platforms. A generalized modular multiplication for NIST primes and its implementation on SIMD platforms is presented. Our results show that accelerating basic cryptographic operations on mobile platforms can have large impact on the performance of secure mobile applications. Our future work includes the system integration of elliptic curve cryptography based protocols on mobile platforms. Since Intel N2800 runs at higher frequency, time taken for scalar multiplication on both platforms is almost same. But it might be interesting to study the power consumption for modular multiplication on both platforms.

ACKNOWLEDGEMENTS

This work was supported in part through a NIST grant 60NANB10D004 and NSF grant 0855095.

REFERENCES

- [1] D. F. Aranha and C. P. L. Gouvêa, “RELIC an Efficient Library for Cryptography”, available online at <http://code.google.com/p/relic-toolkit/>
- [2] ARM Limited, “Cortex-A8 technical reference manual”, revision r3p2, 2010. available online at <http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0344k/index.html>
- [3] D. J. Bernstein and P. Schwabe, “NEON crypto,” in *Proceedings of the 14th international conference on Cryptographic Hardware and Embedded Systems CHES'12*, 2012, pp. 320-339.
- [4] D. J. Bernstein and T. Lange, “eBACS: ECRYPT Benchmarking of Cryptographic Systems”, <http://bench.cr.yp.to>.
- [5] D. J. Bernstein, “Curve25519: new Diffie-Hellman speed records.” in *9th international conference on theory and practice in public-key cryptography, PKC 2006*, New York, NY, USA, 2006, pp. 207-228.
- [6] Dan Garcia, “Great Ideas in Computer Architecture SIMD II,” Oct. 2012. available online at <http://www-inst.eecs.berkeley.edu/~cs61c/fa12/lectures/18LecFa12DLPPII.pdf>
- [7] GNU Multiprecision Library available online at <http://gmplib.org/>
- [8] D. Hankerson, A. Menezes, S. Vanstone, “Guide to Elliptic Curve Cryptography”, Springer 2004.
- [9] Intel, “Evaluation Platforms”, available online at http://www.intel.com/p/en_US/embedded/designcenter/tools/seed-board-program?iid=1032#expired
- [10] Intel, “Intel Atom Processor N2800”, available online at [http://ark.intel.com/products/58917/Intel-Atom-Processor-N2800-\(1M-Cache-1_86GHz\)#infosectionpackagespecifications](http://ark.intel.com/products/58917/Intel-Atom-Processor-N2800-(1M-Cache-1_86GHz)#infosectionpackagespecifications)
- [11] Intel, Using Streaming SIMD Extensions (SSE2) to perform Big Multiplications, available online at http://software.intel.com/sites/default/files/m/c/c/7/4/8/24960-40809_w_big_mul.pdf
- [12] S. Mane, L. Judge and P. Schaumont, “An Integrated Prime-Field ECDLP Hardware Accelerator with High-Performance Modular Arithmetic Units,” in *2011 International Conference on Reconfigurable Computing and FPGAs (ReConFig)*, 2011, pp.198-203.
- [13] S. Morozov, C. Tergino and P. Schaumont, “System integration of Elliptic Curve Cryptography on an OMAP platform,” in *Proceedings of the 9th IEEE Symposium on Application Specific Processors SASP'11*, 2011, pp. 52-57.
- [14] NSA, Mathematical routines for the NIST prime elliptic curves, available online at http://www.nsa.gov/ia/ia_files/nist-routines.pdf
- [15] C. Tergino, “Efficient binary field multiplication on a VLIW DSP”, 2009, Master’s Thesis Virginia Tech - 06222009-150103.
- [16] H. Yan, Z. J. Shi and Y. Fei, “Efficient implementation of elliptic curve cryptography on DSP for underwater sensor networks”, in *Proceedings of the 7th Workshop on Optimizations for DSP and Embedded Systems (ODES-7)*, 2009.
- [17] E. Kasper, “Fast Elliptic Curve Cryptography in OpenSSL”, in *Financial Cryptography and Data Security: FC 2011 Workshops, RLCPS and WECSR*, 2011.