Hyper-Threaded Multiplier for HECC
Gabriel Gallin, Arnaud Tisserand

To cite this version:

HAL Id: hal-01620046
https://hal.archives-ouvertes.fr/hal-01620046
Submitted on 20 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—Modular multiplication is the most costly and common operation in hyper-elliptic curve cryptography. Over prime fields, it uses dependent partial products and reduction steps. These dependencies make FPGA implementations with fully pipelined DSP blocks difficult to optimize. We propose a new multiplier architecture with hyper-threaded capabilities. Several independent multiplications are handled in parallel for efficiently filling the pipeline and overlapping internal latencies by independent computations. It increases the silicon efficiency and leads to a better area / computation time trade-off than current state of the art. We use this hyper-threaded multiplier into small accelerators for hyper-elliptic curve cryptography in embedded systems.

I. INTRODUCTION

Nowadays, numerous applications require strong security levels on small hardware devices. Public-key cryptography (PKC) is mandatory for providing key exchange and digital signature. RSA, the first PKC standard, is too costly for embedded applications, more than 2000-bit keys are currently recommended for an average security level. Elliptic Curve Cryptography (ECC [1]) and Hyper-Elliptic Curve Cryptography (HECC [2]) are known to provide a given security level similar to data dependencies, one cannot feed efficiently this pipeline. For instance, a typical 3-stage pipelined DSP block forces to wait 3 cycles before reusing any result from the DSP block. Due to strong dependencies inside the main loop of partial products generation and modular reduction steps to reduce the width of intermediate data and to gain some speedup. One famous variant is the Coarsely Integrated Operand Scanning (CIOS) method presented by Koc et al. in [5]. However, besides these improvements, Montgomery multiplication still suffers from strong dependencies inside the main loop of partial products accumulation and modular reduction.

Most of modern FPGAs embed many dedicated hardware resources for performing small integer multiplications and accumulations (e.g. 18 × 18=48 bits). They are too small for cryptographic operands but they act as efficient basic blocks for building $F_p$ operators. In order to reach high frequencies, DSP blocks must use several internal pipeline stages. For instance, a typical 3-stage pipelined DSP block forces to wait 3 cycles before reusing any result from the DSP block. Due to data dependencies, one cannot feed efficiently this pipeline during MMM iterations with strong dependencies between iterations. This reduces circuit efficiency with lower utilization of DSP blocks.
Algorithm 1 CIOS algorithm (from [5]).

Require: \( R = 2^r \) such as \( 4 \times P < R, P' = -P^{-1} \pmod{w} \) and \( A, B \) two integers such that \( 0 \leq A, B < 2 \times P \)

Ensure: \( 0 \leq S < 2^w, 0 \leq C < 2^w \).

t \equiv (A \times B \times R^{-1}) \pmod{P}, 0 \leq t < 2 \times P

1: for \( i = 0 \) to \( s - 1 \) do
2: \( t \leftarrow 0 \)
3: \( C \leftarrow 0 \)
4: for \( j = 0 \) to \( s - 1 \) do
5: \((C, S) \leftarrow t_s + C\)
6: \( t_j \leftarrow S\)
7: end for
8: \((C, S) \leftarrow t_0 + q_0 \times P_0\)
9: for \( j = 1 \) to \( s - 1 \) do
10: \((C, S) \leftarrow t_j + q_j \times P_j + C\)
11: \( t_{j-1} \leftarrow S\)
12: end for
13: \((C, S) \leftarrow t_s + C\)
14: \( t_{s-1} \leftarrow S\)
15: \( t_s \leftarrow t_{s+1} + C\)
16: end for
17: return \( t \)

IV. HYPER-THREADED MULTIPLIER

We chose a different approach to improve efficiency in small hardware MM units. We decided to use the classical CIOS method as a basis to design an hyper-threaded modular multiplier (HTMM). Hyper-threading consists in interleaving independent operations in the same hardware resource to fill the pipeline as much as possible. HTMM can be seen as multiple logical multipliers sharing the same physical resource. The CIOS method is described in Algo. 1 where \( P \) is a prime modulus and \( n \) is the operand width. Each operand is split into \( s \) words of \( w \) bits such that \( n = s \times w \).

In our target accelerators, we do not use internal communications of \( n \)-bit words (for \( P \) elements) at each clock cycle. Both operands and results of arithmetic units (including HTMM) are transmitted using \( s \) words of \( w \) bits. Targeted DSP blocks are at most 17-bit wide for one unsigned operand. Using multiples of 17 bits for \( w \) saves area and power since we can avoid internal interfaces to accommodate \( n \)-bit field elements computed on 17-bit DSP blocks (here \( w = 2 \times 17 < n \)). Using more than \( 2 \times 17 \) bits for \( w \) is possible but leads to much larger architectures (this is not interesting for our applications in embedded systems).

CIOS, cf. Algo. 1 performs iterations over words of multiplier operand \( A \). Each iteration corresponds to: partial product of multiplicand \( B \) by word \( A_i \) (Algo.1 lines 4–10); “quotient” \( q_i \) determination from partial product \( A_i \times B \) (Algo.1 line 11); and finally addition of \( A_i \times B \) with product \( q_i \times P \) (Algo.1 lines 12–19) for reduction step. Reduction of partial product is done by discarding least significant word of \( t \). This partial reduced result is added to the next partial product \( A_{i+1} \times B \) during next iteration. Each iteration is thus based on 3 strongly dependent products, and iterations themselves are sequential.

HTMM is based on 3 hardware stages as depicted in Fig. 1. Each one corresponds to one step of the iteration described above. STAGE 1 performs the partial product of one \( w \)-bit word of \( A \) and \( s \) words of \( w \) bits from \( B \) and the addition with the previous reduced partial product from STAGE 3. “Quotient” \( q_i \) is computed in STAGE 2. Finally in STAGE 3, the product \( q_i \times P \) is performed and added to the product \( A_i \times B \) computed in STAGE 1, where the least significant \( w \)-bit word of the result is discarded.

Each stage uses fully pipelined DSP blocks to compute partial products at a high frequency. On our target FPGAs (see Sec. V), using the 3 internal pipeline registers of each DSP block leads to 300 to 400 MHz frequencies instead of about 100 to 200 MHz without full pipeline.

When computing a single MM on a pair of operands \((A^{(0)}, B^{(0)})\), the most significant word at iteration \( i \) will be available after some delay due to internal pipeline registers. Then it cannot be used directly into the next iteration \((i + 1)\). Hyper-threading use the idle stages to compute other independent MMs in parallel. In our
HTMM, one can enter 3 independent pairs of operands \(\{(A^{(0)}, B^{(0)}), (A^{(1)}, B^{(1)}), (A^{(2)}, B^{(2)})\}\) before the first product \(P^{(0)} = A^{(0)} \times B^{(0)}\) is computed. This way, all the stages are full after the very first latency, as illustrated by Fig. 2.

Our HTMM can also receive a new pair of operand (e.g. \(A^{(3)}, B^{(3)}\)) slightly before the oldest product (e.g. \(P^{(0)}\)) is transmitted on the unit output (see right part of Fig. 2).

In the future, we will investigate the use of other numbers of pairs of operands (it seems that 3 is probably the most efficient one for this size of field).

V. Implementation and Comparisons

We implemented 2 HTMM versions on 3 different Xilinx FPGAs: Virtex 5 XC5SLX110T (V5), Virtex 4 XC4VLX100 (V4) and Spartan 6 XC6SLX75 (S6). Version HTMM_BRAM uses dedicated hardware RAM blocks (BRAMs) to store the 3 pairs of operands. Version HTMM_DRAM uses distributed memory in programmable LUTs (lookup tables) of logic slices to store the operands and all intermediate values (no block RAM).

We implemented MM from [6] to help comparisons on the target FPGAs. We obtained very close results for both area and timing: e.g. 37 DSP blocks for 256-bit \(\mathbb{F}_p\) for us and [6].

In order to reduce the iterations number without increasing too much the area, we chose \(w = 34\) bits. In HECC, the currently recommended width for \(\mathbb{F}_p\) elements is at least 128 bits. Then, we set \(s = 4\) words and \(n = 4 \times 34 = 136\) bits. In the future, we will study how to use “rectangular” DSP blocks (e.g. \(17 \times 24\) bits).

Fig. 3 details the architecture of a 128-bit HTMM with configurations of DSP blocks and pipeline registers. Boxes colors correspond to stages of Fig. 1 and Algol. 1. Stages 1/2/3 respectively require 4/3/4 DSP blocks.

HTMM_BRAM, HTMM_DRAM and MM from [6] have been implemented in VHDL using Xilinx CAD tools ISE and iSim 14.7. Tab. I reports implementation results in terms of FPGA resources – number of DSP blocks, BRAMs, FFs (flip-flops), LUTs, slices –, and performances – clock cycles, computation time – for 3 independent MMs. These results correspond to the best place and route (PAR) strategies found after 100 runs of SmartXplorer.

The ECC architecture presented in [6] uses a single MM operator without any BRAM in the unit. But it uses more BRAMs to store large field element in Montgomery domain for Orup optimization (with full width communications in the architecture). In parallel accelerators for HECC, we use small internal communications (\(w\)-bit words) and BRAMs in arithmetic units to store various operands and reduce the bandwidth to the central memory. We then consider them as part of the hardware cost (corresponding values are indicated by a * in Tab. I).

For one single MM, both HTMM versions are less efficient than [6]: 69 clock cycles for HTMM versus 25 for [6]. For 3 independent MMs, HTMM requires 79 cycles against 65 cycles for [6] but with almost half the number of DSP blocks and much higher frequency (due to smaller internal values).

Finally, HTMM leads to faster products for about half the cost of the best state of the art solution (Tab. I). On S6 FPGA, HTMM_BRAM leads to 15% speedup, 48% DSP blocks reduction, 66% BRAMs reduction and 33% slices reduction for 3 independent MMs. HTMM_DRAM provides the same speedup and DSP blocks reduction with only 10% slices reduction but without any BRAM.

VI. HTMM Usage in HECC Based on \(\mu\)Kummer

HECC is considered as an interesting solution for PKC on low-cost processors using solutions based on Kummer surfaces [8]. The recent paper [3] presented \(\mu\)Kummer, a very efficient software implementation on low-cost devices (8-bit AVR and 32-bit ARM microcontrollers), with 30 to 70% speedup for key exchange and digital signature.

We work on the design and prototyping of hardware accelerators for PKC (with several curve based solutions) in FPGAs and ASICs. We implemented the \(\mu\)Kummer HECC solution proposed in [3] on several FPGAs to evaluate how it performs in hardware compared to efficient ECC solutions from the state of the art.

Fig. 4 shows the amount of parallelism in the combined point doubling-addition operation proposed in [3] (a/s are additions/subtractions and M/S are multiplications/squares in field \(\mathbb{F}_p\)).

We designed several HECC hardware accelerators using our HTMM unit and various configurations of the architecture (see [9] for more details). We compared the 256-bit ECC solution from [6] and 2 of our accelerators for \(\mu\)Kummer HECC (named H1 and H2). H1/H2 accelerators have the following configurations:

H1: 1 HTMM, 1 \(\mathbb{F}_p\) adder-subtractor, 1 CSWAP unit (in charge of key management), 1 output register (for results of scalar multiplications), 1 global memory, 1 global control and 1 program memory;

H2: H1 configuration with 2 HTMMs and 2 \(\mathbb{F}_p\) adders-subtractors (instead of 1 unit arithmetic of each type).

One can notice that a HTMM unit handles 3 MMs in parallel (6 with 2 HTMMs). Architecture of accelerator H1 is depicted in Fig. 5 Internal communications are handled by (de)-multiplexors driven by the control program. Accelerator H2 is similar with more arithmetic units.

Both H1 and H2 accelerators for HECC have been fully implemented on V4 and V5 FPGAs for comparison to the ECC solution proposed in [6] (known to be the fastest ECC solution on FPGA without using hundreds of DSP blocks).

The corresponding results are reported in Tab. II. On V5 FPGA, accelerator H1 achieves area reduction by 70% DSP blocks, 30% BRAMs and 49% slices but with a 30% duration increase compared to ECC. Accelerator H2 (with 2 arithmetic units of each type) reduces both duration by 15% and area by 40% DSP blocks, 10% BRAMs and 10% slices.

Our results confirm the interest for HECC compare to ECC: either faster primitives for the same area or half of the area for a similar computation time. HTMM helps to design parallel accelerators without increasing too much the internal
TABLE I

<table>
<thead>
<tr>
<th>Version</th>
<th>FPGA</th>
<th>DSP</th>
<th>BRAM 18K/9K</th>
<th>FF</th>
<th>LUT</th>
<th>Slices</th>
<th>Frequency [MHz]</th>
<th>Clock cycles</th>
<th>Time [ns]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM from [6]</td>
<td>V4</td>
<td>21</td>
<td>0/70</td>
<td>1311</td>
<td>1201</td>
<td>879</td>
<td>252</td>
<td>65</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>21</td>
<td>0/*0</td>
<td>1310</td>
<td>1027</td>
<td>406</td>
<td>296</td>
<td>65</td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>21</td>
<td>0/6/*</td>
<td>1280</td>
<td>1600</td>
<td>540</td>
<td>210</td>
<td>65</td>
<td>309</td>
</tr>
<tr>
<td>HTMM_DRAM</td>
<td>V4</td>
<td>11</td>
<td>0/0</td>
<td>1638</td>
<td>1128</td>
<td>1346</td>
<td>330</td>
<td>79</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>11</td>
<td>0/*0</td>
<td>1616</td>
<td>652</td>
<td>517</td>
<td>400</td>
<td>79</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>11</td>
<td>0/0</td>
<td>1631</td>
<td>1344</td>
<td>483</td>
<td>302</td>
<td>79</td>
<td>261</td>
</tr>
<tr>
<td>HTMM_BRAM</td>
<td>V4</td>
<td>11</td>
<td>2/*0</td>
<td>615</td>
<td>364</td>
<td>449</td>
<td>328</td>
<td>79</td>
<td>241</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>11</td>
<td>2/*0</td>
<td>593</td>
<td>371</td>
<td>249</td>
<td>357</td>
<td>79</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>S6</td>
<td>11</td>
<td>0/2</td>
<td>587</td>
<td>359</td>
<td>180</td>
<td>304</td>
<td>79</td>
<td>260</td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th>FPGA</th>
<th>Accelerator</th>
<th>DSP Blocks</th>
<th>BRAM 18K</th>
<th>Slices</th>
<th>Freq.</th>
<th>Time [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>V4</td>
<td>ECC</td>
<td>37</td>
<td>11</td>
<td>4655</td>
<td>250</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>H1</td>
<td>11</td>
<td>7</td>
<td>1413</td>
<td>330</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>H2</td>
<td>22</td>
<td>9</td>
<td>2536</td>
<td>330</td>
<td>0.35</td>
</tr>
<tr>
<td>V5</td>
<td>ECC</td>
<td>37</td>
<td>10</td>
<td>1725</td>
<td>291</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>H1</td>
<td>11</td>
<td>7</td>
<td>873</td>
<td>360</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>H2</td>
<td>22</td>
<td>9</td>
<td>1542</td>
<td>360</td>
<td>0.32</td>
</tr>
</tbody>
</table>

communications inside the architecture. Using more than 2 HTMM units for \( \mu \)Kummer is currently difficult due to the central data memory.

VII. CONCLUSION

We proposed an hyper-threaded modular multiplier for HECC over 128-bit generic prime fields. It leads to better area / computation time trade-offs than the best state of the art solutions. By improving the hardware efficiency (more hardwired resources are active at each clock cycle), we are able to be slightly faster but a lot smaller than [6] when at 3 independent modular multiplications are computed in parallel.

In the future, we plan to study other HTMM versions. We will also study hyper-threaded schemes impact on energy consumption and on side-channel leakage.

ACKNOWLEDGMENT

This work was partly funded by the HAH project (Labex CominLab and Lebesgue, Brittany Region, [http://www.h-a-h.cominlabs.ueb.eu/](http://www.h-a-h.cominlabs.ueb.eu/)).
Fig. 5. Hardware accelerator for HECC adapted from the \(\mu\)Kummer solution proposed in [3] for software implementation.

REFERENCES