Precision Unwound: Fine-Tuning Loop Unrolling for Energy-efficient FPGA-based PQC using HLS

Srijeet Guha* and Andrea Guerrieri†
*NVIDIA Graphics Private Limited, India
†School of Engineering, HES-SO Valais-Wallis, Switzerland

***Abstract*—High-Level Synthesis (HLS) is a valuable tool for designing hardware accelerators for post-quantum cryptography (PQC). However, while mapping high-level code to hardware, the quality of the synthesized hardware in terms of latency, power, and area are sensitive to various design parameters and configurations, such as loop unrolling, pipelining, and dataflow optimizations. Previous efforts explored the effects of target frequency on the energy consumption of the generated hardware [1]. In this work, we explore the effects of loop unrolling on the execution time and energy efficiency of the final PQC accelerators. We demonstrate that, despite initial expectations, loop unrolling could worsen the performance and quality of designs in certain circumstances.**

I. Introduction

PQC standards were announced by NIST in September 2024. HLS is a valuable tool for designing FPGA-based PQC cores using C/C++. In HLS, the quality of the synthesized hardware in terms of latency, power, and area remains highly sensitive to various design parameters and configurations. Loop unrolling is a common optimization technique in which iterations of a loop are duplicated to allow parallel execution, thus improving performance by reducing latency. This approach can significantly reduce the latency of loop-intensive applications by allowing each loop iteration to operate independently on dedicated hardware resources. Yet, in energy-constrained environments, loop unrolling may be expensive, motivating designers to accurately adjust the unrolling factor.

II. Unrolling Factor versus Energy Efficiency

A classical approach to determining the fastest hardware configuration is to unroll the loop to the maximum extent. However, this is valid in situations where the loop does not involve memory dependencies. However, an increase in the initiation interval (II) of the loop can degrade performance, nullifying the effect of increased parallelism. The relation between the loop unrolling factor and the initiation interval determines the performance and energy efficiency of the final hardware. Our experiments have been performed on the number theoretic transform (NTT) module that executes polynomial multiplication in lattice-based PQC standards. We have used Altera HLS and Quartus Prime v24.01 to compile the optimized NTT code proposed by [2], shown on list 1. We generated multiple designs with different unrolling factors targeting Cyclone-10GX 10CX220YF780I5G FPGA at 100MHz. The results are plotted in Fig. 1.

Listing 1. *NTT*, refactored code from [2].

```
k = 1; zeta = zetas[k++];
for (len = 128; len >= 2; len >>= 1){
   limit = len; start = 0;
   for (int j = 0; j < 128; j++){
   #pragma unroll
      uint16_t r_j_len = r[start + len];
      uint16_t r_j = r[start];
      t = fqmul(zeta, r_j_len);
      r[start + len] = r_j - t;
      r[start] = r_j + t;
      start++;
      if (start == limit){
         start += len;
         limit += (len << 1);
         zeta = zetas[k++]; }}}
```

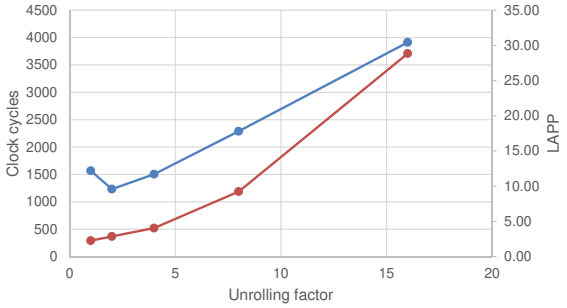

Fig. 1. Unrolling factor versus clock cycles and LAPP (Latency-Area-Power-Product). The results show that a design with minimum clock cycles is achieved with an unrolling factor of 2 (blue line), after which performance decreases linearly. The effects on LAPP are similar, yielding the best results without unrolling.

III. Conclusion and Future Work

In this work, we present a preliminary study on the effects of loop unrolling on performance and the energy efficiency of PQC. Future work includes the study of the analytical methods to predict these effects before synthesis, as well as the trade-offs introduced by different memory configurations.

References

[1] S. Guha and A. Guerrieri, "Iterative frequency tuning targeting energy efficiency ratio for fpga-based post-quantum cryptographic cores," in *2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2024, pp. 1–4.

[2] A. Guerrieri, G. D. S. Marques, F. Regazzoni, and A. Upegui, "Optimizing post-quantum cryptography codes for high-level synthesis," in *2022 Euromicro Conference on digital systems Design (DSD22)*, Gran Canaria, Spain, 2022, pp. 361–67.