ECE 6775 Final Project: FPGA Acceleration of Post-Quantum Cryptography

Aidan McNay (acm289) Barry Lyu (fl327) Edmund Lam (el595) Nita Kattimani (nsk62)

Introduction

With the increasing reliance on digital systems in the modern age, maintaining the security and privacy of important data is more critical than ever. As such, cryptographic algorithms for securly storing and communicating data have widespread usage. To maintain security, these algorithms are centered around computational problems that are infeasible for classical processors to solve; these are known as *NP-hard* problems, which have no known polynomial time solution. Users with a secret key are able to decrypt the encoded messages, but users without would have to solve this NP-hard problem to access the encrypted data, which are designed to take an astronomical number of years to solve with brute-force.

The advent of quantum computers have called the stengths of many of these algorithms into question. For example, both RSA (a popular asymmetric-key algorithm with widespread use) and Diffie-Hellman (an algorithm for securly establishing a common shared key, for use in symmetric-key encryption) rely on the difficulty of factoring large numbers for their security. Algorithms for quantum computers to solve such problems in polynomial time have existed for a while [1], but have never had a computer advanced enough to run them. However, modern advances in quantum computing have demonstrated that computers may be available soon that can crack these algorithms. Even just earlier this week, Google unveiled a new quantum computer, "Willow", that can achieve speedups over the fastest classical processors on select problems by a factor of 10^{30} [2].

While such computers aren't currently able to break modern cryptographic algorithms, many experts suspect it's only a matter of time before current cryptographic algorithms become insecure [3]. To this end, NIST (the National Institute of Standards and Technology) has standardized the use of RSA-2048 only until 2030, and noted that updated strengths are heavily affected by any progress on quantum computing [4]. Additionally, to prepare for the advance of quantum computing, NIST has standardized additional, *quantum-resistant* algorithms [5]. These algorithms are centered around different computational problems for which there is no known algorithm to efficiently solve for both classical and quantum computers. This *post-quantum cryptography* (PQC) will have increasing significance as advances in quantum computers are made. Additionally, since malicious adversaries could already be recording data to later decrypt once sufficient quantum computers are available, PQC algorithms are already being recommended and used for extremely sensitive data, such as government operations [6].

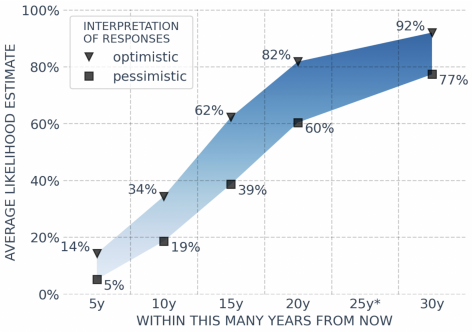


Figure 1: A timeline of when experts believe that RSA-2048 will be able to be broken by a quantum computer in 24 hours [3]

For public-key encryption (where the encrypting and decrypting keys are different), as well as key establishment (for securely establishing a shared secret key over an insecure network), NIST recommends CRYSTALS-Kyber, or simply **Kyber**. Since such an algorithm would be widely used in communication, speeding up its operation would have large impacts for a variety of applications. For our project, we explored implementing Kyber using custom hardware on an FPGA.

Problem Description - Kyber

Components of the Kyber Algorithm

Optimization of NTT

Implementations

FPGA Adaptation

Code Changes

Simulation

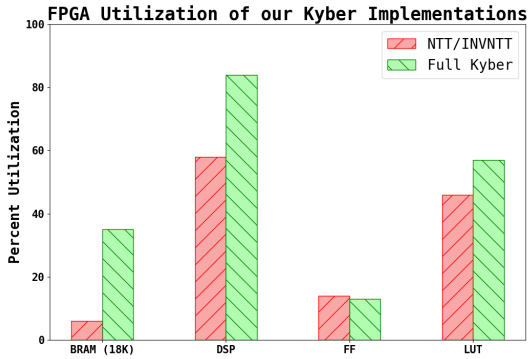
Host Implementation

Evaluation

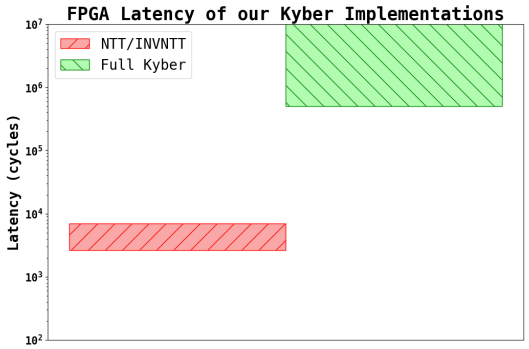
To quantitatively understand our design and the impact of our design choices, we pushed both the implementation of the entire design, as well as the more-optimized NTT kernel (including the NTT and INVNTT operations), through the FPGA flow and implemented on the class Zynq FPGA. This gave us both results after synthesis, as well as a bitstream to use on an FPGA to gain experimental results, allowing us to better gain takeaways about the efficacy of our design.

Synthesis

The first step to gain quantitative results is to use Vivado's HLS flow to push both of our implementations through the FPGA flow. This can tell us how many hardware resources the designs would need, as well as a prediction of how long each design will take. This is shown in Figures 2a and 2b, respectively.



(a) The hardware utilization of the different FPGA implementations



(b) The expected latency of the different FPGA implementations

Figure 2: The synthesis results for our FPGA implementations

For the utilization in Figure 2a, we can see that putting the entire design on FPGA hardware results in more hardware resources overall, as expected. One interesting note is the overall amount of DSP usage; for both designs, a significant number are used, much more than other labs have seen. This is due to the need for multipliers in Montgomery conversion. Having a value in Montgomery form can change performing a modulus over one value (as needed for modular ring operations) to another; by choosing to perform a modulus over a power of two, such moduli can be implemented in hardware quickly. However, converting in and out of this representation involves a multiplication step. This particular operation was one of the key reasons why the NTT kernel couldn't be as optimized. The NTT kernel had an outer loop that required a Montgomery conversion for each iteration, where the number of iterations varied each time it was used. Although we could unroll most uses of this loop, the largest version of this loop could not be unrolled, due to the limited number of DSP blocks available.

While the entire design used more resources in general, we also discovered that the optimized NTT kernel surprisingly used *more* flip-flops (**FF**) compared to the overall design. This seems somewhat surprising, considering that the NTT kernels are included (multiple times) in the entire algorithm, but makes sense once you consider the extra optimizations applied to the NTT algorithms in isolation. Specifically, with just the NTT algorithms, we can exploit loop pipelining and unrolling to achieve much higher performance by doing operations in parallel; however, this comes at the cost of extra flip-flops to store intermediate results in the different pipeline stages, as well as replicated across all loop iterations. This contrasts with the entire design in hardware; due to the lack of available resources, we are able to optimized the design much less, reducing the number of flip-flops needed in this case from the lack of aggressive pipelining/unrolling.

In addition to the hardware utilization of each design, we also compare about the predicted latency of each design; how many cycles each design should take. For the NTT algorithms, the latency is uncertain, as it depends on which operation we're performing. Performing an NTT operation takes 2610 cycles, whereas performing an INVNTT operation takes 4378 cycles due to an extra multiplication reduction step needed at the end.

Experimental

Project Management

Conclusion

Acknowledgements

References

[1] P. Shor, "Algorithms for Quantum Computation: Discrete Logarithms and Factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 124–134, Revised in 1996. [Online]. Available: https://ieeexplore.ieee.org/document/365700

[2] H. Neven, "Meet Willow, our state-of-the-art quantum chip," Dec. 2024. [Online]. Available: https://blog.google/technology/research/google-willow-quantum-chip/

[3] Global Risk Institute, "2024 Quantum Threat Timeline Report," Dec. 2024, accessed December 10th, 2024. [Online]. Available: https://globalriskinstitute.org/publication/2024-quantum-threat-timeline-report/

[4] E. Barker, "Recommendation for Key Management," May 2020. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-57pt1r5

[5] G. Alagic *et al.*, "Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process," July 2022. [Online]. Available: https://doi.org/10.6028/NIST.IR.8413-upd1

[6] "Quantum Computing Cybersecurity Preparedness Act," 44 U.S.C. § 3502, 3552 and 3553 Chapter 35, Dec. 2022. [Online]. Available: https://www.govinfo.gov/app/details/PLAW-117publ260/summary