Accelerating Oblivious Transfer with a Pipelined Architecture

Xiaolin Li[1,2,3], Wei Yan[1,2,3], Hongwei Liu[1,2], Yong Zhang[3], Qinfen Hao[1,2,3], Yong Liu[3], Ninghui Sun[1,2,3]
[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Chinese Academy of Sciences, Beijing, China
[3] Zhongguancun Laboratory, Beijing, China
{lixiaolin19s, yanwei, liuhongwei, haoqinfen, snh}@ict.ac.cn, {zhangyong, liuyong}@zgclab.edu.cn

***Abstract*—With the rapid development of machine learning and big data technologies, ensuring user privacy has become a pressing challenge. Secure multi-party computation offers a solution to this challenge by enabling privacy-preserving computations, but it also incurs significant performance overhead, thus limiting its further application. Our analysis reveals that the oblivious transfer protocol accounts for up to 96.64% of execution time. To address these challenges, we propose POTA, a high-performance pipelined OT hardware acceleration architecture supporting the silent OT protocol. Finally, we implement a POTA prototype on Xilinx VCU129 FPGAs. Experimental results demonstrate that under various network settings, POTA achieves significant speedups, with maximum improvements of $22.67\times$ for OT efficiency and $192.57\times$ for basic operations in MPC applications.**

***Index Terms*—secure multi-party computation, oblivious transfer, pipeline architecture, hardware accelerator**

I. Introduction

Cloud computing has been widely applied in various scenarios such as finance and healthcare. However, with its rapid development, security issues in cloud have become increasingly prominent. Unlike traditional computing scenarios, there is a high probability that user private data may be leaked or inferred by untrusted parties in the process. This issue becomes more prominent with the emergence of machine learning training and inference services in cloud.

Secure multiparty computation (MPC) represents a significant technical approach to address security issues in cloud computing [1]. MPC enables parties to collaboratively compute using their private data without a trusted third party, ensuring data confidentiality while achieving accurate results. However, MPC introduces significant overhead, leading to a drastic performance degradation in cloud applications. For instance, compared to a plaintext multiplication, secret multiplication using MPC protocols may degrade performance by more than five to six orders of magnitude [2].

In this paper, we first present four basic operations widely used in cloud applications, such as multiplication, matrix multiplication, bitwise AND, and comparison. Our study is based on the implementation of MP-SPDZ, the state-of-the-art software MPC framework [3]. We observe that up to 96.64% of the execute time for these operations is consumed by the oblivious transfer (OT) protocols. To overcome these challenges, we propose POTA, a high-performance pipelined hardware architecture for accelerating the silent OT protocol [7], which mitigates the performance bottleneck caused by limited network bandwidth. Moreover, POTA mainly involves two specialized hardware subsystems, the construction of puncturable pseudorandom function (PPRF) and the large matrix-vector multiplication under the learning parity with noise (LPN) assumption. These two phases are the most compute-intensive parts. Ultimately, by enhancing the execution efficiency of OT protocols, the performance of secure computations is improved.

TABLE I
The proportion of execution time of OT in basic operations in various network settings.

Network	Basic Operations			
	Mul	MatMul	AND	CMP
5Gb/s	96.57%	96.57%	96.64%	94.41%
320Mb/s	96.12%	96.13%	96.19%	93.94%
100Mb/s	90.17%	88.63%	90.86%	87.99%

We implement a prototype of POTA on FPGAs and evaluate its performance. The experimental results demonstrate that compared to existing OT works, POTA delivers acceleration ratios of up to $22.67\times$ in various network settings. When compared to current state-of-the-art MPC works in the same network settings, POTA achieves performance improvements of up to $192.57\times$ for basic operations in MPC applications.

II. Analysis of OT Protocols

In this section, we analyze OT protocols for MPC applications in cloud. Our study employs MP-SPDZ [3], the state-of-the-art software MPC framework, including protocols based on OT [4]. We benchmark four basic operations commonly used in MPC applications: multiplication, matrix multiplication, bitwise AND, and comparison. Matrix multiplication involves a single multiplication of two 128×128 matrices, while the other operations are performed one million times each. Our experiments simulate various network settings, including a LAN setting with 5 Gb/s bandwidth and two WAN settings with 320 Mb/s, 100 Mb/s bandwidth, respectively. The results are depicted in Table I. We find that the execution of the OT protocol constitutes up to 96.64% of the total time for the aforementioned four basic operations.

Based on the aforementioned analysis, it is evident that improving the efficiency of the OT protocol can significantly enhance the performance of the MPC applications. However, the efficiency of the OT protocol is constrained by limited network bandwidth. To address this issue, the accelerator supports the silent OT protocol rather than the IKNP protocol, which heavily relies on network bandwidth.

III. The POTA Design

Inspired by the above observation, we present POTA, a pipelined hardware accelerator for MPC applications in cloud. Figure 1 depicts the overall architecture of POTA, comprising two silent OT compute engines for sender and receiver, along with a 100G Ethernet subsystem and PCIe subsystem. One silent OT engine serves as the sender in the OT protocol,

978-3-9826741-0-0/DATE25/© 2025 EDAA

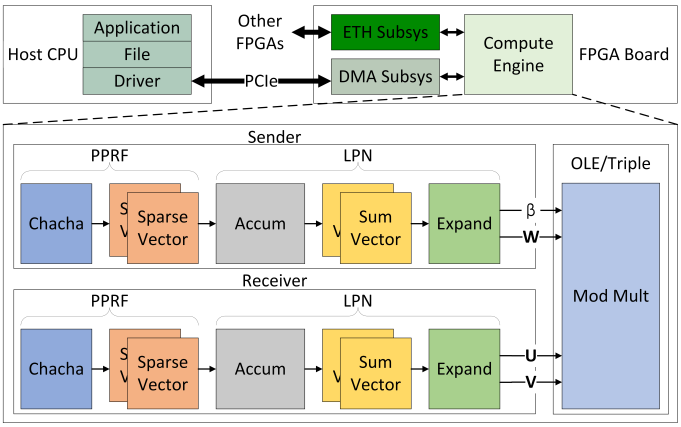

Fig. 1. POTA architecture.

while the other engine as the receiver, allowing users to assume different roles.

The silent OT protocol compute engine consists of two subsystems. The first subsystem is designed for constructing PPRF, effectively mitigating the dual consumption of execution time or hardware resources. The second subsystem implements large-scale matrix-vector multiplication under the LPN assumption, thereby reducing both the utilization of on-chip resources and the required memory bandwidth. PPRF utilizes the ChaCha encryption module as a CSPRNG to generate the GGM tree, which is then saved in Sparse Vector. The Accumulator and Expander modules achieve the large matrix-vector multiplication under the LPN assumption, with intermediate results stored in the Sum Vector. With the outputs from the silent OT protocol compute engine, the OLE/Triple Generator can generate OLE or multiplication triples. The POTA prototype is implemented on FPGAs. Therefore, its 100G subsystem is utilized for network communication among parties, while the PCIe subsystem in the FPGA connects to the host CPU, enabling it to read the generated OT, OLE, and multiplication triples for the MPC applications.

IV. Evaluation

We evaluate the performance of the POTA design for generating OT across three network configurations and OT sizes, and compare it with the IKNP protocol and the silent OT protocol in libOTe framework. The size of OT ranges from 2^{27} to 2^{30}, and the results are presented in Table II. In general,

TABLE II
Comparisons for OT efficiency in various settings.

Network	Size	POTA(s)	IKNP(s)	Silent OT(s)
5Gb/s	2^{27}	4.91	9.30(1.89×)	33.17(6.75×)
	2^{28}	9.83	19.08(1.94×)	76.87(7.82×)
	2^{29}	19.66	36.79(1.87×)	244.82(12.46×)
	2^{30}	39.31	74.26(1.89×)	505.20(12.85×)
320Mb/s	2^{27}	5.85	60.63(10.36×)	38.46(6.57×)
	2^{28}	11.71	121.35(10.37×)	103.64(8.85×)
	2^{29}	23.41	243.42(10.4×)	211.15(9.02×)
	2^{30}	46.81	485.37(10.37×)	526.54(11.25×)
100Mb/s	2^{27}	8.05	182.4(22.66×)	32.01(3.98×)
	2^{28}	16.11	264.44(22.63×)	102.77(6.38×)
	2^{29}	32.21	728.93(22.63×)	214.08(6.65×)
	2^{30}	64.41	1,460.09(**22.67**×)	539.59(**8.38**×)

TABLE III
Comparisons for basic operations in various network settings.

Network	OP	Online(s)	POTA(m)	MP-SPDZ(m)
5Gb/s	Mul	0.12	0.28	18.08(65.46×)
	MatMul	0.25	0.58	37.92(65.44×)
	AND	0.13	0.28	18.09(64.32×)
	CMP	24.63	43.38	2,180.23(50.26×)
320Mb/s	Mul	0.13	0.32	38.81(120.93×)
	MatMul	0.30	0.69	81.41(118.70×)
	AND	0.14	0.33	38.82(118.13×)
	CMP	25.75	48.08	4,667.15(97.07×)
100Mb/s	Mul	0.19	0.45	86.77(**192.57**×)
	MatMul	0.45	0.99	182.01(183.84×)
	AND	0.19	0.45	86.77(192.50×)
	CMP	30.10	60.85	10,418.70(171.22×)

POTA outperforms both IKNP and the silent OT protocol across different network configurations and OT sizes.

Next, we demonstrate how POTA accelerates MPC applications in cloud. OT is primarily used in MPC protocols for generating multiplication triples, essential for secure multiplication during the online phase. We choose MP-SPDZ as the framework for our MPC applications, which implements various MPC protocols. We benchmark against the MASCOT protocol available in MP-SPDZ, which is based on OT. For our benchmarks, we evaluate four basic operations described in Section II.

The results are shown in the Table III. We observe that compared to MP-SPDZ, POTA can provide a performance improvement of 1-2 orders of magnitude, with the highest improvement reaching up to 192.57×. Moreover, with decreasing network bandwidth, POTA can achieve even greater speedups. This demonstrates that POTA can significantly enhance the performance of MPC applications in bandwidth-constrained cloud environments.

Conclusion

In this work, we find that OT protocols account for up to 96.64% of the execution time of MPC applications. To address this challenge, we propose POTA, the first high-performance pipelined OT hardware acceleration architecture. POTA addresses network bandwidth limitations and integrates multiple modules to accelerate various compute-intensive operations. Evaluation of the FPGA prototype shows that under various network settings, POTA provides up to 22.67× for OT efficiency and 192.57× for basic operations in MPC applications.

Acknowledgment

This work is supported by National Key R&D Program of China (2022YFB4401501) and The Jiangsu key R&D plan project (BE2023006-4).

References

[1] Jiang H, Xu Q L. Secure multiparty computation in cloud computing[J]. Journal of computer research and development, 2016.
[2] Zhou X, Xu Z, Wang C, et al. PPMLAC: high performance chipset architecture for secure multi-party computation. In ISCA, 2022.
[3] Keller M. MP-SPDZ: A versatile framework for multi-party computation. In CCS, 2020.
[4] Keller M, Orsini E, Scholl P. MASCOT: faster malicious arithmetic secure computation with oblivious transfer. In CCS, 2016.
[5] Boyle E, Couteau G, Gilboa N, et al. Efficient two-round OT extension and silent non-interactive secure computation. In CCS, 2019.
[6] Reisert P, Rivinius M, Krips T, et al. Overdrive LowGear 2.0: Reduced-Bandwidth MPC without Sacrifice. In Asia CCS, 2023.
[7] Boyle E, Couteau G, Gilboa N, et al. Efficient pseudorandom correlation generators: Silent OT extension and more. In CRYPTO, 2019.
[8] Boyle E, Couteau G, Gilboa N, et al. Correlated pseudorandomness from expand-accumulate codes. In CRYPTO, 2022.