

FLAIRS: FPGA-Accelerated Inference-Resistant & Secure Federated Learning

Li, Huimin; Rieger, Phillip ; Zeitouni, Shaza ; Picek, Stjepan; Sadeghi, Ahmad Reza

DOI

[10.1109/FPL60245.2023.00046](https://doi.org/10.1109/FPL60245.2023.00046)

Publication date

2023

Document Version

Final published version

Published in

Proceedings of the 2023 33rd International Conference on Field-Programmable Logic and Applications (FPL)

Citation (APA)

Li, H., Rieger, P., Zeitouni, S., Picek, S., & Sadeghi, A. R. (2023). FLAIRS: FPGA-Accelerated Inference-Resistant & Secure Federated Learning. In L. O'Conner (Ed.), *Proceedings of the 2023 33rd International Conference on Field-Programmable Logic and Applications (FPL)* (pp. 271-276). IEEE.
<https://doi.org/10.1109/FPL60245.2023.00046>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

FLAIRS: FPGA-Accelerated Inference-Resistant & Secure Federated Learning

Huimin Li*, Phillip Rieger[†], Shaza Zeitouni[†], Stjepan Picek^{‡*} and Ahmad-Reza Sadeghi[†]

*Delft University of Technology, The Netherlands, H.Li-7@tudelft.nl

[†]Technische Universität Darmstadt, Germany, {phillip.rieger, shaza.zeitouni, ahmad.sadeghi}@trust.tu-darmstadt.de

[‡]Radboud University, The Netherlands, stjegan@computer.org

Abstract—Federated Learning (FL) has become very popular since it enables clients to train a joint model collaboratively without sharing their private data. However, FL has been shown to be susceptible to backdoor and inference attacks. While in the former, the adversary injects manipulated updates into the aggregation process; the latter leverages clients' local models to deduce their private data. Contemporary solutions to address the security concerns of FL are either impractical for real-world deployment due to high-performance overheads or are tailored towards addressing specific threats, for instance, privacy-preserving aggregation or backdoor defenses.

Given these limitations, our research delves into the advantages of harnessing the FPGA-based computing paradigm to overcome performance bottlenecks of software-only solutions while mitigating backdoor and inference attacks. We utilize FPGA-based enclaves to address inference attacks during the aggregation process of FL. We adopt an advanced backdoor-aware aggregation algorithm on the FPGA to counter backdoor attacks. We implemented and evaluated our method on Xilinx VMK-180, yielding a significant speed-up of around 300 times on the IoT-Traffic dataset and more than 506 times on the CIFAR-10 dataset.

Index Terms—FPGA Acceleration, Federated Learning (FL), FPGA-based FL, Backdoor-aware FL, Privacy-preserving FL

I. INTRODUCTION

FPGAs are powerful and versatile devices, providing flexible platforms for custom hardware solutions. With unique characteristics like parallel processing, support for various data types, low latency, and lower power consumption compared to general-purpose computing platforms, they excel in accelerating computations and tackling complex challenges. FPGAs have become indispensable across various domains, from high-performance computing and data centers to the Internet of Things (IoT) and embedded systems. Their widespread adoption is evident in their deployment within commercial cloud platforms such as Amazon EC2 [1], Microsoft Azure [2], and Alibaba Cloud [3], underscoring their significance and impact in today's technological landscape.

In addition to their inherent benefits, FPGAs enable establishing a trusted execution environment (TEE), ensuring the security of critical workloads, including the FPGA configuration, which may comprise an Intellectual Property (IP) design, and the processed data, without compromising performance. Recent advancements in FPGA research have demonstrated the feasibility of establishing TEEs on commodity FPGAs deployed in cloud environments [4], [5]. Consequently, FPGAs

can provide not only acceleration but also secure processing of clients' workloads in hostile cloud environments. This paradigm shift toward trusted execution on cloud FPGAs offers numerous advantages. It grants organizations greater control over their applications and data security, even when physical access to the FPGA is limited or non-existent. We refer to TEEs on FPGAs as FPGA-based TEEs to distinguish them from TEEs on CPUs.

Federated Learning. One of the compelling applications for FPGA-based TEEs is Federated Learning (FL), a collaborative learning approach. Unlike traditional centralized learning, FL allows clients to train their own DNN models locally using their private datasets and share only the training results with a central server that aggregates clients' models or local models into a global model and returns it to the clients. Hence, sensitive data remains confined to clients' computing premises. Therefore, FL is foreseen to improve the clients' privacy [6].

However, FL is still prone to privacy attacks during the aggregation process. Such attacks aim to infer information about the training data of a model, e.g., if a specific sample was used [7] or try to reconstruct samples from the training data [8]. Although individual clients' contributions are anonymous, preventing associating the inferred information with a specific client, a malicious aggregation server can still exploit access to the local models to analyze them and violate clients' privacy.

Another type of attack on FL targets the model's integrity. Poisoning attacks aim to manipulate the global model to misbehave, i.e., *targeted or backdoor attacks* [9], [10], [11], or they aim at rendering the model useless, i.e., *untargeted attacks* [12], [13]. Targeted attacks are crucial because the adversary injects a stealthy function to influence the outcome without violating the model's utility.

Existing Defenses. Proposed defenses for FL typically address only one type of attack, focusing either on protecting clients' privacy [14], [15], [16] against a potentially malicious server, or mitigating specific backdoor attacks launched by malicious clients [10], [17], [18], [19]. Mitigating both types of attacks poses a dilemma. On the one hand, detecting and filtering poisoned models requires the aggregator to inspect local models. On the other hand, privacy defenses prevent the aggregator from inspecting the local models. This presents a challenge in striking the right balance between security and privacy in FL presents a challenge.

To solve this dilemma, several privacy-preserving ap-

proaches such as Homomorphic Encryption (HE) or Secure Multi-Party Computation (SMPC) [20], [21], [22] have been proposed to process models without divulging any information. However, such solutions result in significant performance overhead and scalability issues, particularly for complex backdoor defenses involving vector metric computations or clustering. Implementing these defenses using SMPC, such as in the case of [21], [20], becomes highly impractical due to the associated overhead. An alternative approach is to utilize TEEs on CPUs to ensure local models' privacy while the aggregator inspects them. For instance, TEE-based implementations of Krum [17] have been explored [23], [24]. However, TEEs' limited computation capacities introduce significant overhead for computation-intensive algorithms like Krum, which involves calculating Euclidean distances between local models.

Therefore, utilizing FPGA-based TEEs seems to be an intuitive approach for achieving secure and privacy-preserving FL. Among recent software-based proposals, FLAME aims to address backdoor and inference attacks to be independent of the attack strategy. To counter backdoor attacks, FLAME combines outlier-detection-based filtering with model clipping and noising. However, FLAME suffers from significant performance overhead due to the deployment of SMPC for protecting clients' privacy. SMPC protocols enable the secure evaluation of a public function, e.g., the aggregation process, on private data, e.g., local models, from N mutually distrusting parties. SMPC finds utility in outsourcing scenarios [25], where multiple parties/clients can secret-share their private inputs among two or more non-colluding, well-connected, and powerful servers responsible for executing the SMPC protocol, yet resulting in a significant computation overhead and hence does not scale. In the case of FLAME, for aggregating 50 models trained on the CIFAR-10 dataset, SMPC increases the execution time of FLAME from 2.6s to 766.1s. Additionally, SMPC requires non-colluding aggregation servers. Consequently, the privacy guarantees of FLAME only apply to semi-honest aggregation servers that adhere to the SMPC protocol [21], [25].

Goals and Contributions. In this work, we propose to leverage FPGA-based TEEs to enable privacy-preserving backdoor-aware aggregation for FL. Our optimized FPGA-accelerated approach enables the aggregation server to perform a privacy-preserving backdoor analysis of the local models with only a negligible computation overhead. The described techniques allow the acceleration of arbitrary backdoor defenses. We exemplify our approach using the recently proposed defense FLAME [21].

Our contributions can be summarized as follows:

- We leverage FPGA-based TEEs, demonstrating a practical and efficient backdoor-aware FL aggregation while protecting clients' privacy in a stronger adversary model than SMPC.
- Our approach generally allows implementing arbitrary backdoor resilient aggregation schemes on secure FPGAs.
- We demonstrate the security and performance gain of FLAIRS by realizing exemplarily the entire FLAME [21]

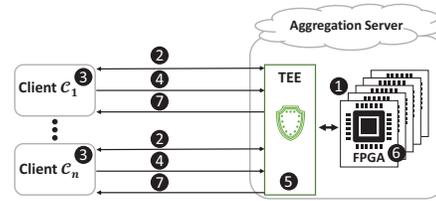


Fig. 1: Workflow of FLAIRS.

algorithm on FPGA, resulting in speed-ups of over 288 times on the IoT-Traffic dataset and more than 506 times on the CIFAR-10 dataset. The above results are obtained from a single FPGA. However, the runtime can be reduced by almost m if m FPGAs run in parallel.

- We propose the cascade structure, enabling the calculation of cosine distance with time complexity of $O(n)$ instead of $O(n^2)$. This structure also significantly reduces the access time to the main memory.

II. BACKGROUND

Remote Attestation verifies the authenticity and integrity of code or memory on a remote device. The *verifier* receives a signed cryptographic hash of the content from a trusted component on the *prover* device. The received digest is compared against the expected reference value to determine the prover's status.

Trusted Execution Environments (TEEs) provide secure enclave applications isolated from all other untrusted software in the system. The TEEs aim to protect the confidentiality and integrity of the enclave's code and data. Remote attestation can be used by clients before sharing confidential data with enclaves. Code and data are processed unencrypted in the CPU's caches and registers but are encrypted and integrity-protected with an enclave-specific secret key before storing in untrusted storage. Commercial TEEs like Intel SGX [26], AMD SEV [27], and ARM TrustZone [28] are widely deployed in computing systems, including FPGA-SoCs such as ARM TrustZone on Intel Stratix 10 SoC and AMD Xilinx ZCU102.

TEEs on FPGAs. FPGA trusted execution protects IP configuration and processed data. FPGA manufacturers like Intel and AMD Xilinx offer hardened cryptographic cores for IP confidentiality and integrity. For cloud deployments without physical access, a trusted third party or vendor is necessary [29], [30], [31], [4]. Thereafter, two methods of establishing FPGA TEEs exist (1) loading IP design on SoCs with built-in TEEs to guarantee FPGA trustworthiness [32], [5]; and (2) using a trust anchor (secure shell) for FPGAs without TEEs [4] which provides remote key generation, configuration, isolation, and cryptographic operations. Here, remote attestation ensures the authenticity and integrity of the FPGA configuration [33], [4].

III. PROBLEM SETTING

A. System & Adversary Model

We consider a system consisting of an aggregation server and n clients jointly training a DNN model. Each client C_i

trains locally a model with its dataset and sends the aggregator its local model. The aggregation runs on the cloud and can therefore benefit from one or multiple FPGAs to accelerate the aggregation. The system setting is visualized in Fig. 1, together with the individual steps of FLAIRS, which will be described in Sect. III-B.

We consider two types of adversaries: (1) \mathcal{A} that aims to inject a backdoor, and (2) \mathcal{A}^S that aims to infer information about clients' training data.

To inject a backdoor, \mathcal{A} changes the predictions of all samples within a trigger set $\mathcal{I} \subset \mathcal{D}$ towards a specific label. \mathcal{A} must ensure that the attack is not detected, which includes preventing a significant drop in the aggregated model's utility on regular samples. We assume \mathcal{A} fully controls $n_{\mathcal{A}} < n/2$ clients and can manipulate their training process and data. However, \mathcal{A} does not know the data or models of other clients.

On the malicious server side, aligned with existing work, \mathcal{A}^S aims to extract information from the individual local models [34], [21], [20] since the aggregation anonymizes the individual contributions of clients and prevents an adversary from associating information gained from the aggregated model with particular clients. We assume that \mathcal{A}^S controls the aggregation server and a few clients, has full software-level access, and can arbitrarily deviate from the aggregation process. We exclude denial-of-service attacks intended to shut down computational resources, as they can be detected. Moreover, we assume that physical attacks on the infrastructure, including the FPGAs, are out of scope. However, remote physical attacks performed using malicious FPGA configurations can be mitigated using FPGA scanners [35], [36], as demonstrated in [4]. As we show next, all clients can vet the FPGA configurations that represent the accelerators and verify their integrity and authenticity as a part of attesting the TEE.

B. FLAIRS Overview

To achieve secure and practical backdoor-aware FL, we adapt FLAME [21] to run on a FPGA-based TEE [29], [30], [31], [4], [32], [5]. Note that the entire aggregation algorithm (demonstrated in Sect. IV-A) is unlikely to fit on a single FPGA, considering a large number of clients and model parameters. Therefore, the aggregation algorithm can be split into several accelerators and benefit from using several FPGAs or swapping in and out the different accelerators on a single FPGA. Hence, when multiple FPGAs/accelerators are used, a scheduler algorithm must coordinate the work of the accelerators and receive clients' models. The scheduler can be implemented as a software application in a TEE or a hardware IP continuously running on the FPGA. In both cases, the scheduler can be attested by clients to ensure its authenticity and integrity. In the following, we describe FLAIRS's workflow (Fig. 1).

Step ①. This step establishes a TEE on the cloud FPGA, where clients' models can be processed securely [4], [5].

Step ②. The clients attest the TEE, i.e., verify the integrity and authenticity of the FPGA configurations that process clients' models. Thus, the clients have the assurance that the

code processing their models is benign, i.e., not corrupted by \mathcal{A}^S , and can exchange secret session keys with the TEE to encrypt their models.

Step ③. The clients encrypt their models using individual secret session keys exchanged with the TEE.

Step ④. The clients send their encrypted models to the aggregator.

Step ⑤. The models are then stored in memory, ready for aggregation.

Step ⑥. The TEE and FPGAs use FLAME to aggregate the models and mitigate backdoor attacks.

Step ⑦. The aggregated model is sent back to all clients.

IV. DESIGN & IMPLEMENTATION

A. Analysis of FLAME Algorithm

We adopt FLAME [21] for backdoor-aware aggregation, which consists of three defense layers, namely Model Filtering, Model Clipping, and Noising, shown in Fig. 2. Adding noise can remove the backdoor but will also drop the models' utility in terms of accuracy on the main task. While filtering and clipping decrease the amount of noise required to mitigate poisoned models. We thoroughly analyze FLAME and define efficient hardware components or processing elements (PEs).

To optimize FLAME performance, we break down the compute-intensive cosine distance (Model Filtering) into two parallel components: preprocessor and cosine similarity. **Preprocessor** component *Prep. PE* calculates the differential vector and Euclidean distances L_2_norms between local models and the global model. **Cosine-similarity** component *CosinePE* runs parallel to *Prep. PE* and computes the dot products of clients' differential vectors, which, along with L_2_norms , are used to calculate cosine distances for all local models. **HDBSCAN** (Model Filtering) labels local models based on cosine distances as benign models 1 or malicious models 0. **Scale** (Model Clipping) calculates the median value (S_t) of L_2_norms and generates scaled models for all local models. HDBSCAN and Scale PEs run in parallel with no data dependency. The **Aggregation** *Agg. PE* comprises clipping (Model Clipping), aggregation, and noise addition (Noising) steps. It generates the final aggregated model by using models' labels, median values, and model scales from previous components.

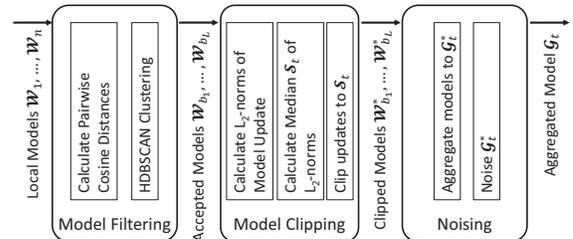


Fig. 2: High-Level Overview of FLAME [21]

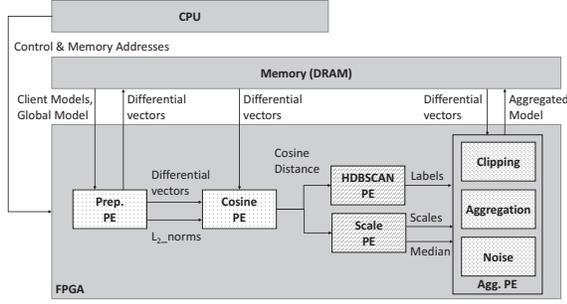


Fig. 3: System Architecture of FLAIRS

B. Implementation

We implemented FLAIRS (Fig. 3) on Xilinx Vitis 2022.2 using the VMK180 Evaluation Kit, which has a built-in ARM TrustZone in its hardened processing unit. The system comprises a host program running on a TEE-enabled CPU and an FPGA platform consisting of shell and kernel components. The host program manages the FPGA’s components and the operation process [37]. The `shell` provides the essential functions for execution, security, and communication interfaces, while the `kernel` is the dynamic region for custom logic implementation. We utilized HLS to translate the C++ kernel module into device logic fabric and RAM/DSP block [38]. The operating frequency of the `kernel` is set to 300 MHz, and a burst mode was adopted to utilize AXI4 interface throughput fully. The width of the AXI4 Memory Mapped interface was configured to 512 bits, and the burst length is set to transfer 4KB each time [38].

1) *Preprocessor PE*: The Preprocessor PE stores the global model locally on the FPGA, followed by the sequential transmission of local models from DDR-RAM to FPGA. Each local model undergoes subtraction with the global model to obtain its differential vector, which is then routed to the *CosinePE* and back to DDR-RAM. After that, each differential value is squared and accumulated, and the accumulated value is processed using the square root function to yield the L_2_norms value for each client. This process operates parallel within a pipeline structure defined by the total number of clients and parameters per local model.

2) *Cosine-similarity PE*: The pairwise cosine distances can be represented as a matrix. The values on the diagonal are equal to 0, while the remaining positions are determined by Eq. (1), where d denotes the differential vector and p represents the total number of parameters. The denominator of Eq. (1) is obtained from L_2_norms , and the numerator is derived from the dot product of two differential vectors of clients i and j . The cosine distance matrix is symmetric, so we only compute the upper triangular or lower part.

$$dist_{ij} = 1 - \frac{d_i d_j}{\|d_i\| \|d_j\|} = 1 - \frac{\sum_{k=1}^p d_i^k d_j^k}{\sqrt{\sum_{k=1}^p (d_i^k)^2} \sqrt{\sum_{k=1}^p (d_j^k)^2}} \quad (1)$$

We use a cascade structure (Fig. 4b) to obtain dot products with multiple stages, where each stage locally stores the first-arriving differential vector in RAM. The cosine distance is calculated for the entire row using this initial vector and

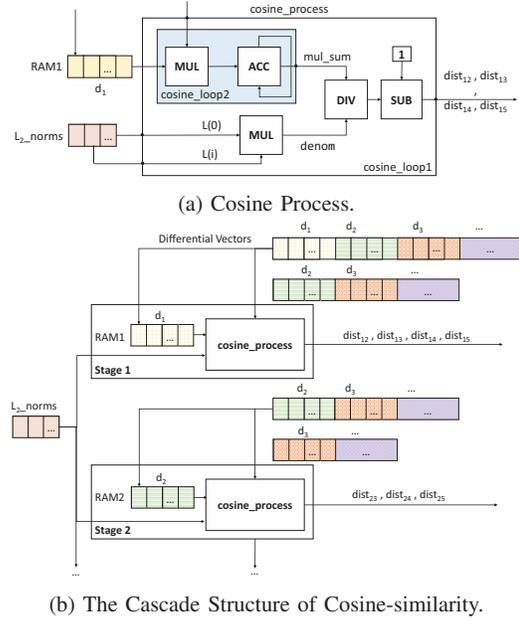


Fig. 4: Cosine-similarity PE.

subsequent differential vectors in the *cosine_process* phase (Fig. 4a). Each stage, except the last one, sends out differential vectors (excluding the first vector) to its subsequent stage. The total number of stages depends on the device resources, number of clients, and number of parameters. If the number of stages is insufficient to calculate cosine distance for all local models, we utilize `hls :: burst_maxi<>` to manually read the remaining differential vectors in bursts from DDR-RAM after each operation of the cascade structure. The new differential vectors are then processed through the same cascade structure for further cosine distance computations.

3) *HDBSCAN PE*: HDBSCAN PE determines one cluster representing the majority of models, thus containing at least $n/2 + 1$, the minimum number of benign local models. The remaining models not part of this cluster are then considered noise. For clustering, FLAME uses HDBSCAN based on the implementation of McInnes *et al.* [39]. Based on the FLAME’s parametrization of HDBSCAN, a simplified version is implemented in the HDBSCAN PE, neglecting all unused aspects. For example, there is no need to determine whether two closely packed dense groups of models constitute separate clusters or a single cluster.

4) *Scale PE*: Here, we first sort L_2_norm values from smallest to largest and then select the median value (S_i) from the sorted list. Subsequently, we calculate $\gamma[i] = \frac{S_i}{L_2_norms[i]}$, where i represents the client order. Finally, we obtain the model scale for each client as $\min(1, \gamma[i])$.

5) *Aggregation PE*: In the aggregation PE, locally stored differential vectors from cascade structures are sequentially multiplied by corresponding scale values and added to the global model for models labeled as 1 to get the resultant value *add_sum*. If RAM does not have enough space to store all dif-

TABLE I: Runtime in seconds of FLAME using FLAIRS (F) compared to FLAME using SMPC (S) for n clients.

| Dataset | n | Cosine Distance | | HDBSCAN | | Scale | | Aggregation (+ Clipping +Noise) | | Kernel Runtime | | Data Transfer (Host↔ DDR) | | Runtime | | Speed-up |
|-------------|-----|-------------------------|-------------------------|---------|-------------------------|-------------------------|-------------------------|------------------------------------|-------------------------|-------------------|--------|------------------------------|---|---------|--|----------|
| | | F | S | F | S | F | S | F | S | F | S | F | S | | | |
| IoT-Traffic | 10 | 5.3553×10^{-2} | 1.5628×10^{-5} | 3.64 | 1.420×10^{-6} | 1.6465×10^{-2} | 7.0034×10^{-2} | 1.0677×10^{-2} | 8.0711×10^{-2} | 108.16 | 1340.1 | | | | | |
| | 50 | 0.453 | 1.509×10^{-3} | 41.84 | 1.4019×10^{-5} | 0.238 | 0.693 | 1.1809×10^{-2} | 0.705 | 269.35 | 382.1 | | | | | |
| | 100 | 2.072 | 1.1851×10^{-2} | 253.87 | 5.2265×10^{-3} | 0.939 | 3.023 | 1.2245×10^{-2} | 3.035 | 876.96 | 288.9 | | | | | |
| CIFAR-10 | 10 | 0.21 | 1.5628×10^{-5} | 3.64 | 1.420×10^{-6} | 4.0984×10^{-2} | 0.251 | 1.2104×10^{-2} | 0.263 | 134.93 | 513 | | | | | |
| | 50 | 1.235 | 1.509×10^{-3} | 41.84 | 1.4019×10^{-5} | 0.263 | 1.5 | 1.2258×10^{-2} | 1.512 | 766.12 | 506.7 | | | | | |

ferential vectors, we use `hls :: burst_maxi<>` to retrieve the remaining vectors from DDR-RAM. `add_sum` is accumulated for all local models, which is then divided by `accepted_num` (number of benign local models with label 1) to generate a quotient. This quotient is added to the noise created using the `MT19937IcnRng` function from Xilinx’s Vitis Library [40], generating random numbers following a normal distribution $N(0, 1)$. The output of `MT19937IcnRng` is multiplied with the required range λ to conform to FLAME’s noise range.

6) *The Scheduler*: The scheduler is the host program that runs on the TEE-enabled CPU and orchestrates the work of the FPGA accelerators, i.e., the kernels. Once the aggregation process is initialized, the scheduler is set as an enclave application. Clients attest the authenticity and integrity of the scheduler to ensure its code has not been modified. The scheduler then receives encrypted local models, decrypts & re-encrypts them with a unified secret key, and stores them in the DDR-RAM. The scheduler detects the Xilinx device, attests the FPGA binary file, and programs it into the device. Then, it creates the buffers the kernel needs in the DDR-RAM and sets up the kernel’s input parameters (mapping ports). Later, the scheduler writes data into buffers, triggers kernel execution, and waits for notification upon completion. Finally, it reads and sends the aggregated model to clients.

C. Evaluation

To ease the comparison with FLAME, we replicated the setup in [21] and evaluated FLAIRS on two datasets: IoT-Traffic and CIFAR-10, with varying values of n of 10, 50, and 100. For the FLAME algorithm, we show the runtime of each component and overall system in Tab I. Note that the runtime of FLAIRS includes the time taken by the Kernel and data transfer between the host and the FPGA. For the IoT-Traffic dataset, we have achieved significant performance enhancement of 1 340.1, 382.1, and 288.9 times with n being 10, 50, and 100, respectively. On the CIFAR-10 dataset, we obtained a speed-up of 513 and 506.7 for n values of 10 and 50, respectively. Compared to the implementation without SMPC, FLAME requires approximately 2.62 seconds for 50 CIFAR10 models, whereas our approach only takes about 1.5 seconds, underscoring the superiority of using FPGAs for accelerating operations.

Our evaluation has demonstrated the impressive ability of our accelerators to effectively enhance performance across different datasets and varying values of n . The results mentioned above are obtained from a single FPGA. However, it is worth emphasizing that the computation of both Cosine Distance and Aggregation can be partitioned into multiple FPGAs and

calculated simultaneously, thereby reducing latency. These two components can account for up to 98% of the total runtime. Therefore, running m FPGAs simultaneously can lead to a runtime reduction of almost m . This distributed FPGA computing approach offers a promising solution to further enhance the performance of our proposed framework.

V. RELATED WORK

Privacy-Preserving Backdoor Defenses in FL. Baffle et al. [34] requires clients to inspect wrong predictions of the aggregated model, making it compatible with SMPC. However, attackers can avoid detection by not changing not-triggered sample predictions. Trusted hardware approaches, such as poisoning defenses on TEEs, have been proposed but are impractical due to significant overhead [24], [23]. SMPC has been used for various approaches against poisoning attacks but has limited scalability due to expensive operations [21], [20], [41].

FPGA Acceleration for FL. Previous research studies have used FPGAs to speed up FL with HE for client privacy like the HE framework presented in [42] and the HW/SW co-design utilized in [16]. However, neither of these works addresses backdoor attacks. Currently, no work has explored using FPGAs for SMPC acceleration.

VI. CONCLUSIONS

In this paper, we have presented FLAIRS, a framework that capitalizes on the benefits of FPGA-based computing to overcome performance bottlenecks inherent in software-only solutions. We demonstrate how FPGA-based TEEs can be leveraged to enable practical and privacy-preserving backdoor-aware FL aggregation on cloud FPGAs. FLAIRS offers more robust security guarantees than SMPC while minimizing the performance overhead. Its flexibility allows for the implementation of arbitrary aggregation schemes on secure FPGAs. Furthermore, our successful FPGA-accelerated implementation demonstrates the exceptional computational capabilities of FPGAs for accelerating FL algorithms.

Acknowledgment. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1119 – 236615297, HMWK within the ATHENE project, the Hessian Ministry of Interior and Sport within the F-LION project, Intel as part of the Private AI Center, and Huawei as part of the OpenS3 Lab.

REFERENCES

- [1] Amazon AWS, “Amazon EC2 F1,” <https://aws.amazon.com/ec2/instance-types/f1/>.

- [2] Microsoft Research, "Project Catapult," <https://www.microsoft.com/en-us/research/project/project-catapult/>.
- [3] A. Cloud, "FPGA-based Compute-Optimized Instance Families," <https://www.alibabacloud.com/help/doc-detail/108504.htm>, 2019.
- [4] S. Zeitouni, J. Vliegen, T. Frassetto, D. Koch, A.-R. Sadeghi, and N. Mentens, "Trusted configuration in cloud fpgas," in *International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2021.
- [5] M. Zhao, M. Gao, and C. Kozyrakis, "Shef: Shielded enclaves for cloud fpgas," in *ACM SPLOS*. ACM, 2022.
- [6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, 2017.
- [7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *IEEE S&P*, 2017.
- [8] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, e, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *USENIX Security*, 2020.
- [9] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in *AISTATS*, 2020.
- [10] S. Shen, S. Tople, and P. Saxena, "Auror: Defending Against Poisoning Attacks in Collaborative Deep Learning Systems," in *ACSAC*, 2016.
- [11] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning Attacks on Federated Learning-Based IoT Intrusion Detection System," in *Workshop on Decentralized IoT Systems and Security*, 2020.
- [12] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.
- [13] M. Fang, X. Cao, J. Jia, and N. Gong, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," in *USENIX Security*, 2020.
- [14] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *CCS*, 2017.
- [15] H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, H. Möllering, T. D. Nguyen, P. Rieger, A.-R. Sadeghi, T. Schneider, H. Yalame *et al.*, "Safelearn: Secure aggregation for private federated learning," in *IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021.
- [16] Z. Wang, B. Che, L. Guo, Y. Du, Y. Chen, J. Zhao, and W. He, "Pipefl: Hardware/software co-design of an fpga accelerator for federated learning," *IEEE Access*, vol. 10, pp. 98 649–98 661, 2022.
- [17] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *NIPS*, 2017.
- [18] N. M. Jebreel and J. Domingo-Ferrer, "Fl-defender: Combating targeted attacks in federated learning," *Knowledge-Based Systems*, vol. 260, p. 110178, 2023.
- [19] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection," in *NDSS*, 2022.
- [20] Y. Khazbak, T. Tan, and G. Cao, "Mlgard: Mitigating poisoning attacks in privacy preserving distributed collaborative learning," in *International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2020.
- [21] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A.-R. Sadeghi, and T. Schneider, "FLAME: Taming backdoors in federated learning," in *USENIX Security*. USENIX Association, 2022.
- [22] Y. Tian, R. Wang, Y. Qiao, E. Panaousis, and K. Liang, "Flvoogd: Robust and privacy preserving federated learning," *arXiv preprint arXiv:2207.00428*, 2022.
- [23] A. Mondal, Y. More, R. H. Rooparagunath, and D. Gupta, "Poster: Flatee: Federated learning across trusted execution environments," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 707–709.
- [24] H. Hashemi, Y. Wang, C. Guo, and M. Annavaram, "Byzantine-robust and privacy-preserving framework for fedml," in *ICLR Workshops*, 2021.
- [25] D. Demmler, T. Schneider, and M. Zohner, "Aby-a framework for efficient mixed-protocol secure two-party computation," in *NDSS*, 2015.
- [26] Intel, "Intel software guard extensions," <https://software.intel.com/content/www/us/en/develop/topics/software-guard-extensions.html>.
- [27] D. Kaplan, J. Powell, and T. Woller, "Amd memory encryption," https://developer.amd.com/wordpress/media/2013/12/AMD_Memory_Encryption_Whitepaper_v7-Public.pdf, 2016.
- [28] ARM, "ARM TrustZone technology," <https://developer.arm.com/ip-products/security-ip/trustzone>.
- [29] K. Eguro and R. Venkatesan, "FPGAs for Trusted Cloud Computing," in *IEEE International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2012, pp. 280–287.
- [30] B. Hong, H.-Y. Kim, M. Kim, T. Suh, L. Xu, and W. Shi, "Fasten: An fpga-based secure system for big data processing," *IEEE Design & Test*, 2017.
- [31] M. E. Elrabaa, M. Al-Asli, and M. Abu-Amara, "Secure Computing Enclaves Using FPGAs," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2019.
- [32] N. Khan, S. Nitzsche, A. G. López, and J. Becker, "Utilizing and extending trusted execution environment in heterogeneous socs for a pay-per-device ip licensing scheme," *IEEE TIFS*, vol. 16, pp. 2548–2563, 2021.
- [33] J. Vliegen, M. M. Rabbani, M. Conti, and N. Mentens, "Sacha: Self-attestation of configurable hardware," in *DATE*, 2019.
- [34] S. Andreina, G. A. Marson, H. Möllering, and G. Karame, "BaFFLe: Backdoor Detection via Feedback-based Federated Learning," in *ICDCS*, 2021.
- [35] J. Krautter, D. R. Gnad, and M. B. Tahoori, "Mitigating Electrical-level Attacks Towards Secure Multi-Tenant FPGAs in the Cloud," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, 2019.
- [36] T. La, K. Mätas, N. Grunchevski, K. Pham, and D. Koch, "FP-GADefender: Malicious Self-Oscillator Scanning for Xilinx UltraScale+ FPGAs," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, 2020.
- [37] AMD Xilinx, Inc., "Vitis Unified Software Platform Documentation," <https://docs.xilinx.com/r/en-US/ug1393-vitis-application-acceleration>, 2022.
- [38] —, "Vitis High-Level Synthesis User Guide," <https://docs.xilinx.com/r/en-US/ug1399-vitis-hls>, 2022.
- [39] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.
- [40] AMD Xilinx, Inc., "Vitis accelerated-libraries," https://github.com/Xilinx/Vitis_Libraries, 2022.
- [41] Y. Dong, X. Chen, K. Li, D. Wang, and S. Zeng, "Flod: Oblivious defender for private byzantine-robust federated learning with dishonest-majority," in *European Symposium on Research in Computer Security*. Springer, 2021.
- [42] Z. Yang, S. Hu, and K. Chen, "Fpga-based hardware accelerator of homomorphic encryption for efficient federated learning," *arXiv preprint arXiv:2007.10560*, 2020.