Towards Accelerated Healthcare Federated System Through Heterogeneous Accelerators

Giuseppe Sorrentino
Politecnico di Milano
giuseppe.sorrentino@polimi.it

Davide Conficconi
Politecnico di Milano
davide.conficconi@polimi.it

***Abstract*—Federated Learning (FL) enables collaborative model training across multiple hosts without exposing sensitive data. Yet, privacy-preserving is critical, introducing a non-negligible overhead. Furthermore, while GPUs perfectly suit AI workloads, FPGAs outperform them in networking and cryptographic tasks as Fully Homomorphic Encryption (FHE), widely employed in FL to safeguard privacy. In this context, applications like Image Registration (IR), requiring different compute-intensive steps, each suitable for a different architecture, even worsen the hardware dichotomy. Thus, this research explores innovative federated network structures to reduce privacy risks while assigning roles to each host, guaranteeing complete hardware acceleration per each compute-intensive task by partitioning each step on different hardware layers of modern heterogeneous platforms as Neural Processing Unitss (NPUs), combining AI Engines (AIEs) and GPUs, Versal VCK5000, combining FPGA and AIEs, and multi-platform peer-to-peer systems.**

I. Problem & Motivation

Federated Learning (FL) enables distributed deep learning by keeping data local, which enhances privacy. In FL, nodes send model updates to a central aggregator, but privacy measures—like homomorphic encryption—add latency and demand hardware acceleration. While GPUs are excellent for deep learning, they often fall short for cryptographic or networking tasks, whereas FPGAs excel due to their reconfigurability. This hardware mismatch becomes even more critical in applications such as Image Registration (IR), which require FPGA-accelerated pre-processing alongside GPU-driven deep learning. Motivated by these challenges, our research explores modern AI accelerators that combine diverse hardware layers, such as AMD NPUs, combining AI Engines (AIEs) and GPUs, Versal VCK5000, combining FPGA and AIEs, and multi-platform peer-to-peer systems. By partitioning the network and assigning roles to each peer, we can split FL workload properly across the multiple peers, focusing on the acceleration of different compute-intensive steps per each role.

II. Background & Related Works

Federated Learning- In federated learning, clients train their local model to compute weights and share model information with a central aggregator (Figure 1 - *First Level Federated Network*). This communication is secured through privacy-preserving mechanisms such as noise addition, blockchains or the more promising FHE [1]. Central aggregators are critical. Indeed, they combine the weights of multiple local models to produce a single global model shared with all the peers. By doing so, each peer benefits from others' information without breaking privacy, ensuring accurate results even for models trained with insufficient data [2]. Given its importance, the literature explores multiple aggregation techniques, distributing the workload, to avoid centrality, or combining such information to ensure a general model improvement [3], [4].

Image Registration- It aligns a 2D or 3D image, namely *floating*, against an accurate reference to correct the distortion caused in the acquisition phase [5]. The procedure consists of a heuristic-based pre-processing step to correct rigid deformations (Figure 1 - top left) and a Deep Learning (DL) phase (Figure 1 - bottom left) to correct non-rigid distortions. Given the complexity, literature accelerates each step, but while DL suits GPUs perfectly, the pre-processing, which is the actual application bottleneck, would benefit from GPUs for image transformation [6], [7] and FPGAs for similarity metric computation [8]. Given this dichotomy, no hardware-accelerated solution perfectly accomplishes each task.

III. Approach & Uniqueness

Accelerating healthcare FL presents several challenges, motivating our novel federated network structure illustrated in Figure 1 (right-side). In our approach, edge nodes focus on IR, handling both inference and local training and forwarding local model weights to a single aggregator. Since these updates are infrequent, edge nodes do not need to accelerate privacy-preserving mechanisms. Furthermore, as we split the whole network in multiple *First Level Federated Networks* and only aggregated models exit the sub-network, specific host model information is kept more private.

In contrast, the aggregator collects local model updates from multiple neighbouring edge nodes, aggregates them, and returns an updated global model to its first-level network and all other aggregators across the *Second Level Federated Network*. By doing so, the information about the sub-network global model is shared with the other sub-networks without disclosing local information about a single peer. Given its multiple connections and diverse incoming data, the aggregator

* DOI: 10.1109/FPL68686.2025.00055

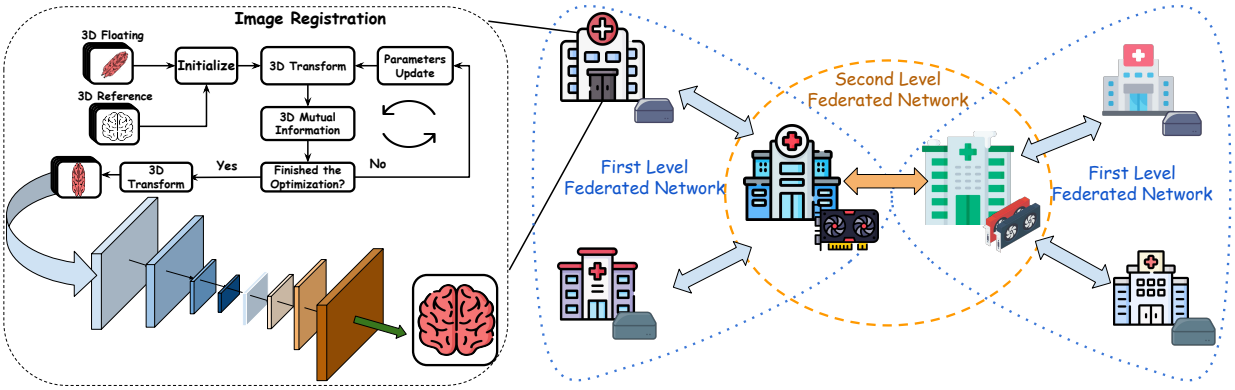

Fig. 1. Multi-Level Healthcare FL Network. Local Aggregators communicate each-other to share information between the two sub-network, while each sub-network trains its own aggregated model. The aggregator leverages one or multiple accelerator cards to enhance both application and communication, while edge nodes can rely on heterogeneous system with NPU, CPU and GPU for application and local model training.

relies on hardware acceleration for the privacy-preserving mechanism—a process that would otherwise be a significant bottleneck. Notably, the aggregator does not train a local model to avoid additional overhead but benefits from the global model. In practical scenarios, such as in surgical ones, using the model is crucial, even without training a local one.

We are currently targeting Ryzen AI for edge nodes, using NPUs and GPUs for efficient training and IR, and Versal systems for aggregators, combining AIEs for IR with FPGAs for fast privacy-preserving operations and networking.

We first target the common application stage. Our **first milestone** focuses on accelerating the most compute-intensive part of IR: the pre-processing step (Figure 1 - top left). Versal accelerator cards are ideal here, using AIEs for image transformations and FPGAs for Mutual Information (MI) computations [9]. A similar approach applies to edge nodes, compensating for the lack of FPGA with optimized CPU or GPU kernels for MI [10]. The **second milestone** integrates pre-processing with the IR inference step (Figure 1 - bottom left). This integration requires careful hardware partitioning, as inference may run on AIEs or FPGAs/GPUs based on the target hardware and performance needs. The **third milestone** targets accelerating FHE, crucial for FL. We plan to offload this task to the FPGA component in aggregators while CPUs handle it on edge nodes.

IV. Preliminary Results

Our preliminary results focus on the pre-processing step of IR (**first milestone**), where our heterogeneous Versal system achieves a speedup of $1.79\times$ compared to leading GPU-based solutions using NVIDIA A100 while being $4.96\times$ more energy efficient. These findings confirm that heterogeneous architectures enable innovative partitioning strategies, enhancing both performance and energy efficiency relative to traditional GPU or FPGA solutions.

V. Expected Results & Contributions

We envision that our FL strategy will lead to measurable improvements across key performance indicators within the complete IR pipeline, directly addressing the rigorous demands of healthcare applications. Considering the IR pre-processing on the heterogeneous versal systems, we have already attained remarkable improvements in both performance and efficiency, representing an important milestone in rigid registration acceleration. Then, by leveraging a federated network structure, we effectively assign specific roles and computational tasks, allowing to have multiple nodes, each accelerating suitably its most compute-intensive tasks. This will minimize federated overhead while ensuring accurate results for peers who lack enough data.

Acknowledgment

This work has financial support from ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU. We thank the AMD Fund for Academic Research and University Program for their support.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[2] M. J. Sheller, B. Edwards, G. A. Reina *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, 2020.

[3] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," 2020.

[4] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," 2019.

[5] A. A. Goshtasby, *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications*. USA: Wiley-Interscience, 2005.

[6] S. Asano, T. Maruyama, and Y. Yamaguchi, "Performance comparison of fpga, gpu and cpu in image processing," *2009 International Conference on Field Programmable Logic and Applications*, pp. 126–131, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:968572

[7] G. Sorrentino, M. Venere, E. D'Arnese, D. Conficconi, I. Poles, and M. D. Santambrogio, "Athena: a gpu-based framework for biomedical 3d rigid image registration," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2023, pp. 1–5.

[8] G. Sorrentino, M. Venere, D. Conficconi, E. D'Arnese, and M. D. Santambrogio, "Hephaestus: Codesigning and automating 3d image registration on reconfigurable architectures," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 5s, pp. 1–24, 2023.

[9] G. Sorrentino, P. S. Galfano, E. D'Arnese, and D. Conficconi, "Soaring with trilli: an hw/sw heterogeneous accelerator for multi-modal image registration," in *33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2025.

[10] A. Zeni, E. Del Sozzo, E. D'Arnese, D. Conficconi, and M. D. Santambrogio, "Starlight: A kernel optimizer for gpu processing," *J. Parallel Distrib. Comput.*, vol. 187, no. C, May 2024. [Online]. Available: https://doi.org/10.1016/j.jpdc.2023.104832