FDL 2025 Work-in-Progress Paper

Efficient Privacy-Preserving Federated Learning on Edge with Reconfigurable FPGA

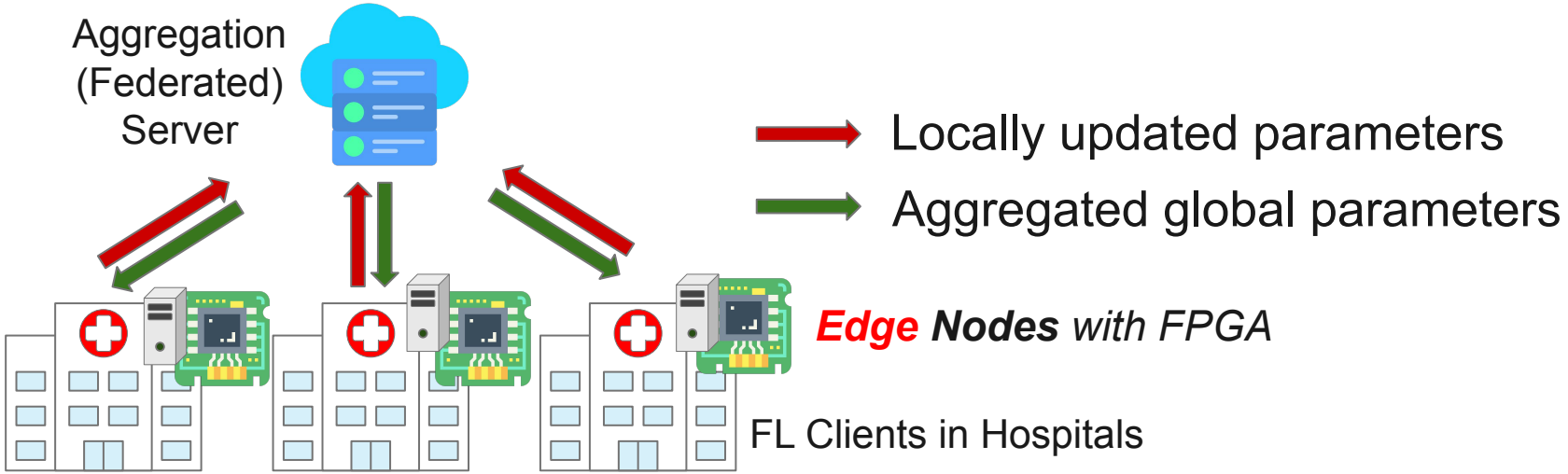
Forum on specification & Design Languages

Sep 10, 2025, at St. Goar, Germany

Byeonggil Jun, Megan Kuo, Aditya A. Krishnan, and Hokeun Kim

Introduction

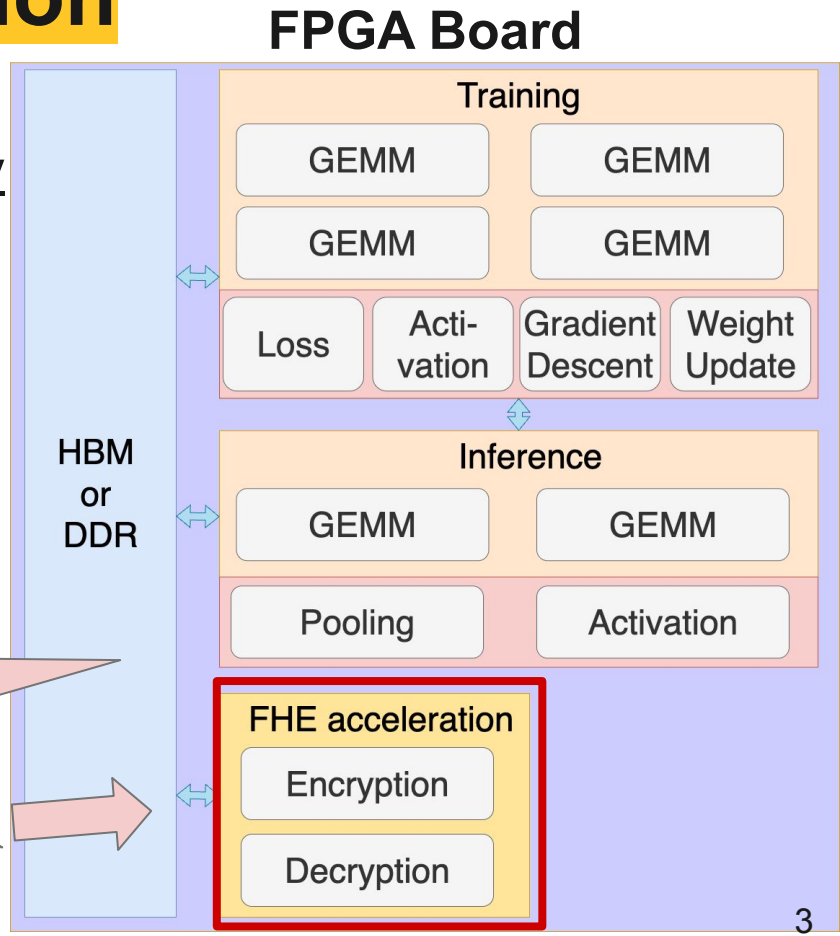
- **Federated Learning (FL)** is often integrated with **Fully Homomorphic Encryption (FHE)**[1] to Enhance Privacy
- However, **FHE with FL** has excessive comp/mem overhead
- **Our WiP Solution**: **FPGA**-based acceleration of FHE+FL



[1] Gong, Y., Chang, X., Mišić, J. et al. Practical solutions in fully homomorphic encryption: a survey analyzing existing acceleration methods. Cybersecurity 7, 5 (2024).

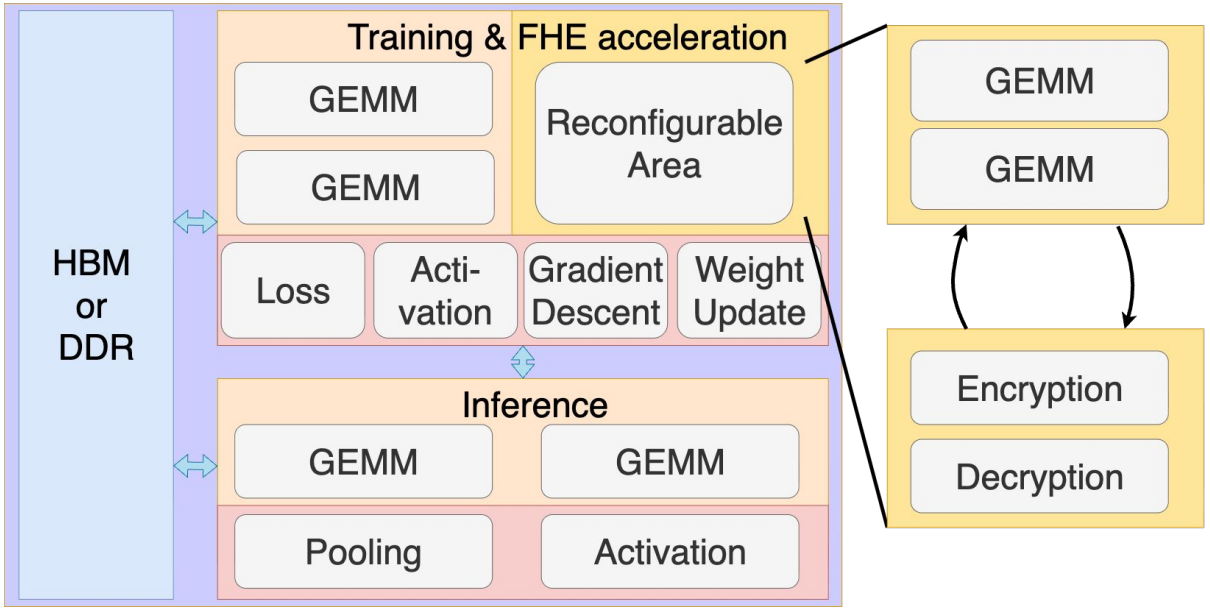
System Model & Motivation

- Resource-constrained clients on the Edge use FPGAs with limited capacity
- *Occasional* FHE operations vs. *Continuous* ML Workload
- **Dynamic Partial Reconfiguration (DPR)** for efficiently using FPGA resources



Work-in-Progress Approach

- Replace some general matrix multiplication (GEMM) modules with encryption/decryption modules for parameter aggregation (WiP)
- Ensure the availability of inference



For ML Workload (Inference & Training)

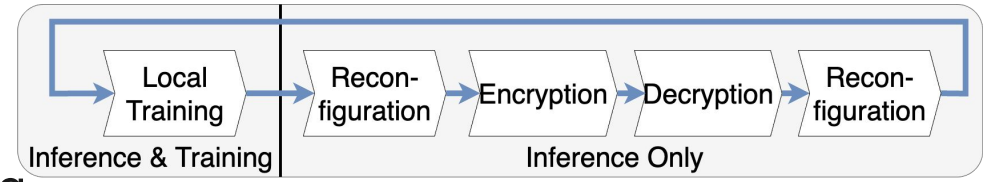
For FL Model Updates + ML Workload

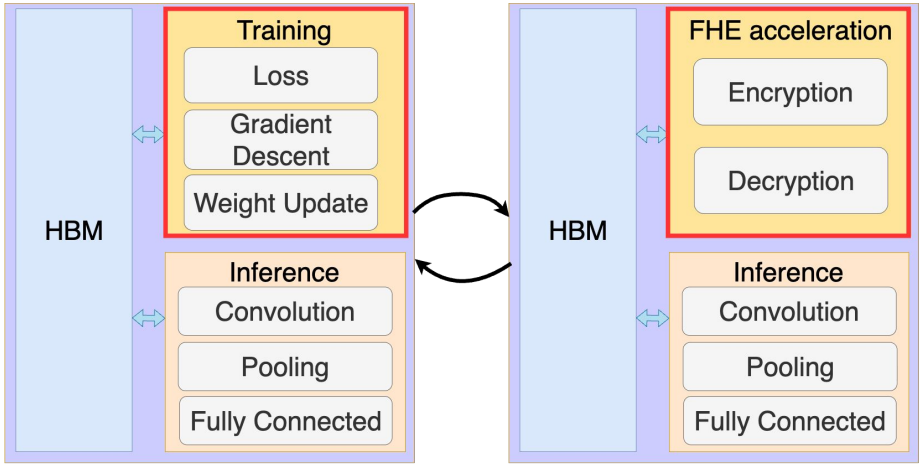
Challenges

- Coordinating FL with dynamic reconfiguration in FPGA
  - E.g., FL update scheduling, Variable number of GEMM modules
- Optimization of trade-off between ML vs. FHE performance
  - E.g., Replacing GEMM modules slows down the training speed
- Utilizing interconnect for reconfigurable vs. non-reconfigurable areas
  - E.g., Interconnect must be static - how can we maximize the interconnect utilization?

Proof-of-Concept Implementation

- Use BGV[2], an FHE algorithm, and LeNet-5, a simple CNN
- Reconfigure the area for training for FL model updates (aggregating parameters)
- Use the full reconfiguration instead of partial reconfiguration (For now)





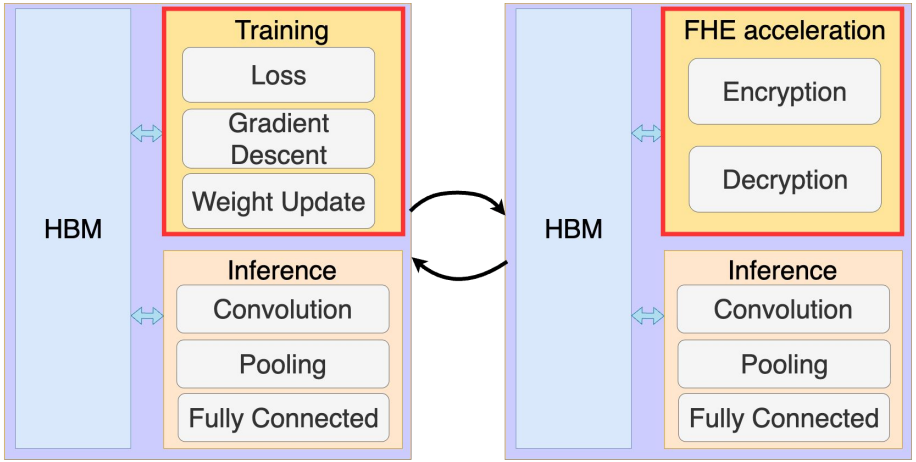
a) An FPGA configuration with inference and training modules.

b) An FPGA configuration with inference and FHE modules.

[2] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. 2014. (Leveled) Fully Homomorphic Encryption without Bootstrapping. ACM Trans. Comput. Theory 6, 3, Article 13 (July 2014), 36 pages.

Preliminary Evaluation of PoC Implementation

- Evaluated on AMD/Xilinx U55C FPGA



a) An FPGA configuration with inference and training modules.

b) An FPGA configuration with inference and FHE modules.

Task on FPGA	LUTs	Latency
FHE Modules	108,179	1.79 ms
Inference	48,961	1.89 ms
Training	92,649	4.03 ms

Task	Latency
Full Reconfiguration	3,912.00 ms

Conclusion & Future Work

- A WiP efficient PoC FPGA-accelerated FL client system using reconfiguration
- Future work toward DPR of the PoC by addressing the remaining challenges
  - Design the accelerator using GEMM modules
  - Examine the trade-off between ML vs. FHE performance
- Extension to consider the design for the servers, especially for decentralized federated learning[3] (any node can perform model aggregation)

[3] E. T. Martínez Beltrán et al., "Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges," in IEEE Communications Surveys & Tutorials, vol. 25, no. 4, pp. 2983-3013, Fourthquarter 2023

- **Contact**
  - byeonggil@asu.edu (***Byeonggil Jun***: Lead Author)
  - hokeun@asu.edu (***Hokeun Kim***: Presenter)
  - https://labs.engineering.asu.edu/kim/ (**ASU KIM Lab**)

Thank you!

- **PoC Implementation**
  - https://github.com/asu-kim/fpga-fl-bgv

ASU KIM
KNOWLEDGEABLE & INTERACTIVE MACHINES