Accelerating Differential Privacy-Based Federated Learning Systems

Mirco Mannino, Alessio Medaglini, Biagio Peccerillo, and Sandro Bartolini

UNIVERSITÀ DI SIENA 1240

Department of Information Engineering and Mathematics
University of Siena, Italy

e-mail: mannino@diism.unisi.it

IARIA

18th International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP) 2024
Special Session: Hardware Accelerators and Accelerated Programming (HAAP)

29th September, 2024

ABOUT ME

- PhD student at University of Siena, Italy
- Research interests:
  - Computer Architecture
  - Architectural Simulation
  - Hardware Accelerators
  - Virtual Memory
  - Deep Learning
- Work Experience:
  - CPU Architect Intern (Huawei R&D, Cambridge UK)
  - SW Embedded Developer (AidiLAB, Siena, IT)

Motivation and Background

Training Using Data Collected on-the-edge

CONVENTIONAL APPROACH

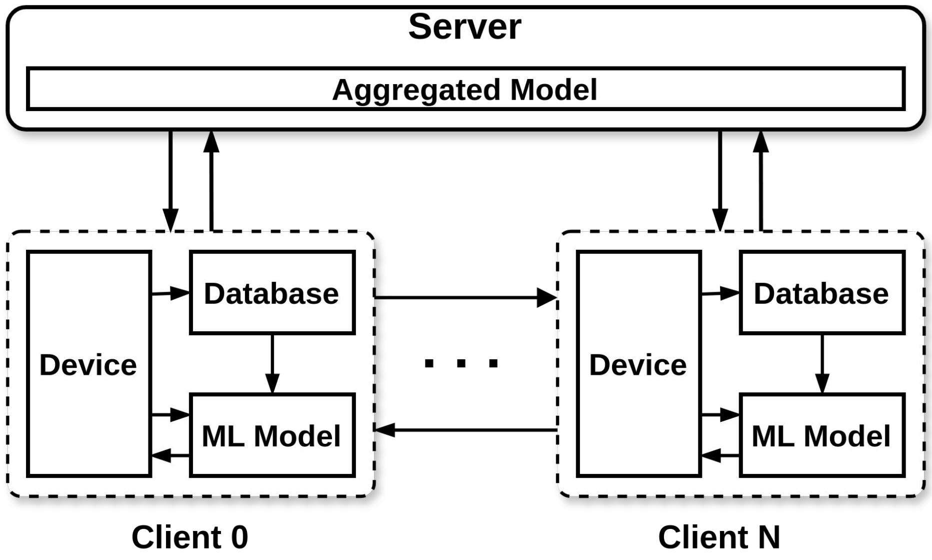
1. Collect data on edge devices (e.g., smartphone)
2. Send data to a central server
3. Train machine learning models in the server
4. Share the trained model to edge devices

It can lead to several **disadvantages**:

- Performance degradation
- Lack of privacy

Federated Learning (1)

- Federated Learning (FL) was Introduced by Google in 2017
- It is a distributed training approach



MAIN STEPS

1. The server shares an untrained model among clients
2. Each client performs a local training procedure using its own data
3. Clients send trained models to the central server
4. The server aggregates them into an updated model
5. The server shares the updated model among clients.

Federated Learning (2)

ADVANTAGES

- User data privacy protection
- Improved model accuracy and diversity
- Bandwidth efficiency

DISADVANTAGES

- Implementation Complexity
- (Possible) Missing HW resources in edge devices

Federated Learning (3)

HOW TO ENSURE USER PRIVACY?

- A key aspect in FL is ensuring the privacy of data collected locally
- Differential Privacy (DP) is one of the promising approaches to ensure user data privacy

DIFFERENTIAL PRIVACY

- Add noise to either data or model guarantee privacy
- Popular approaches:
  - Local Differential Privacy techniques
  - Differential Privacy-based distributed Stochastic Gradient Descent
  - Differential Privacy meta learning

Federated Learning (4)

HARDWARE RESOURCES

- Usually, edge devices are thought for inference, not training
- Acceleration can be achieved in different ways:
  - Graphics Processing Unit (GPU)
  - Field Programmable Gate Array (FPGA)
  - Application Specific Integrated Circuit (ASIC)
- Devices need to be efficiently readapted to meet training needs

Federated Learning Processing Unit

(Possible ideas)

New Challenges

An efficient implementation of Differential Privacy-based Federated Learning system requires:

- **Robust framework** allowing the orchestration of all the players in the system
- **Algorithmic improvement for Differential Privacy,** both on client and server side
- **Specialized hardware** in heterogeneous architectures to accelerate common operations, ensuring energy efficiency

New Challenges

An efficient implementation of Differential Privacy-based Federated Learning system requires:

- **Robust framework** allowing the orchestration of all the players in the system
- **Algorithmic im** client and server side
- **Specialized har** celerate common operations, ens

Several open source solutions:

- FATE
- Federated Tensorflow

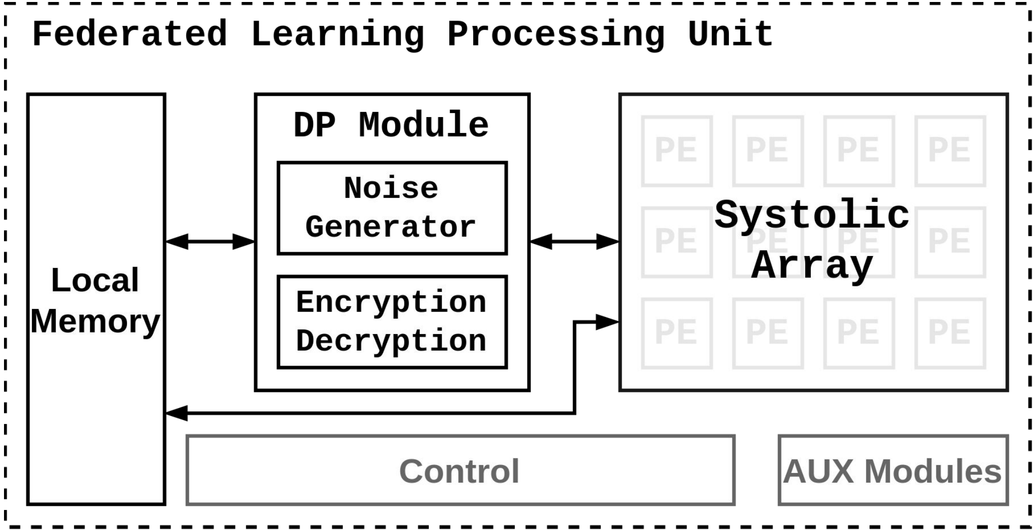
New Challenges

An efficient implementation of Differential Privacy-based Federated Learning system requires:

- **Robust framework** allowing the orchestration of all the players in the system
- **Algorithmic improvement for Differential Privacy,** both on client and server side
- **Specialized hardware** in heterogeneous architectures to accelerate common operations, ensuring energy efficiency

Federated Learning Processing Unit

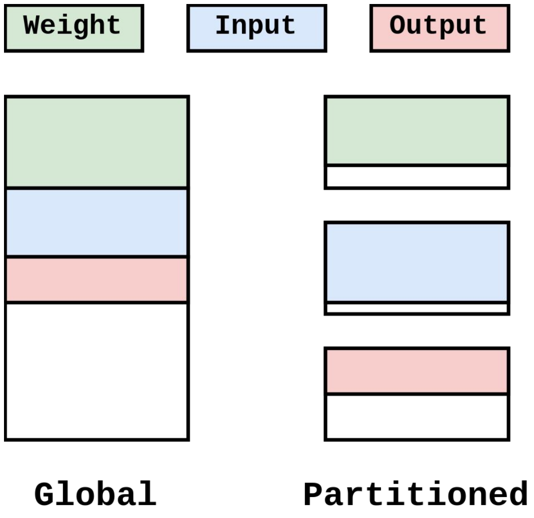
- Dedicated hardware module to speed up DP-based FL systems
- Its implementation requires analysis from several points of view

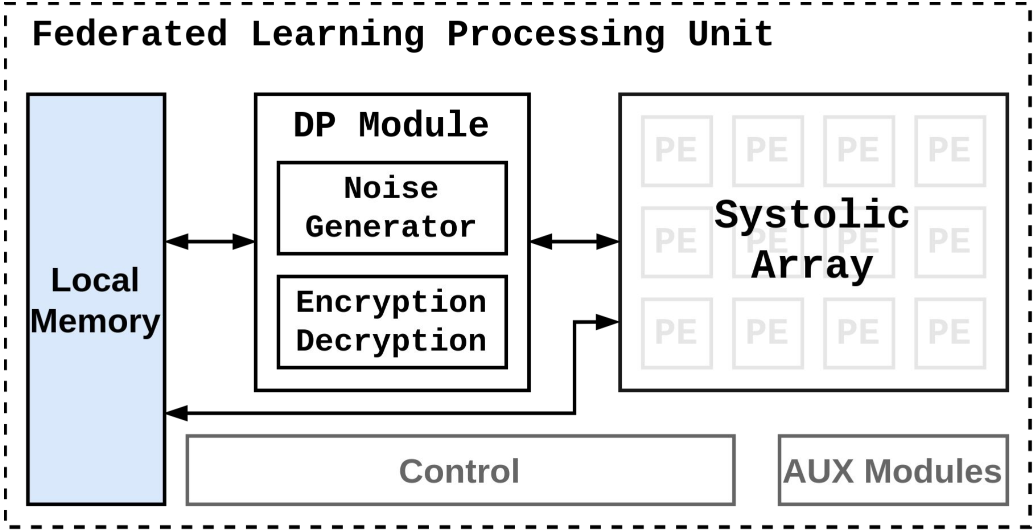


Federated Learning Processing Unit

Possible design choices for **local memory**:

- **Global** buffer
- **Partitioned** local memories





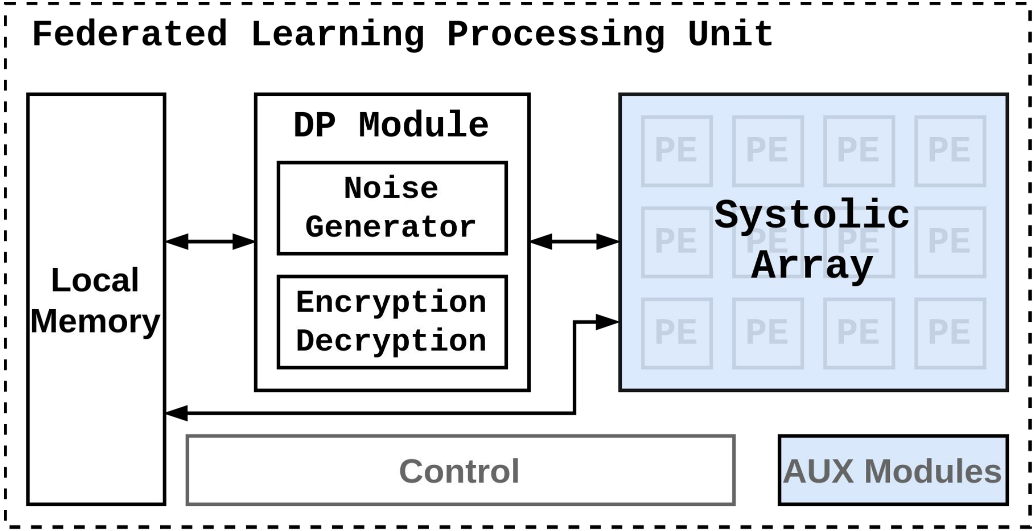
Federated Learning Processing Unit

Possible design choices for **systolic array**:

- Output stationary dataflow
- Weight stationary dataflow
- Input stationary dataflow

Auxiliary modules:

- Activation
- Quantization

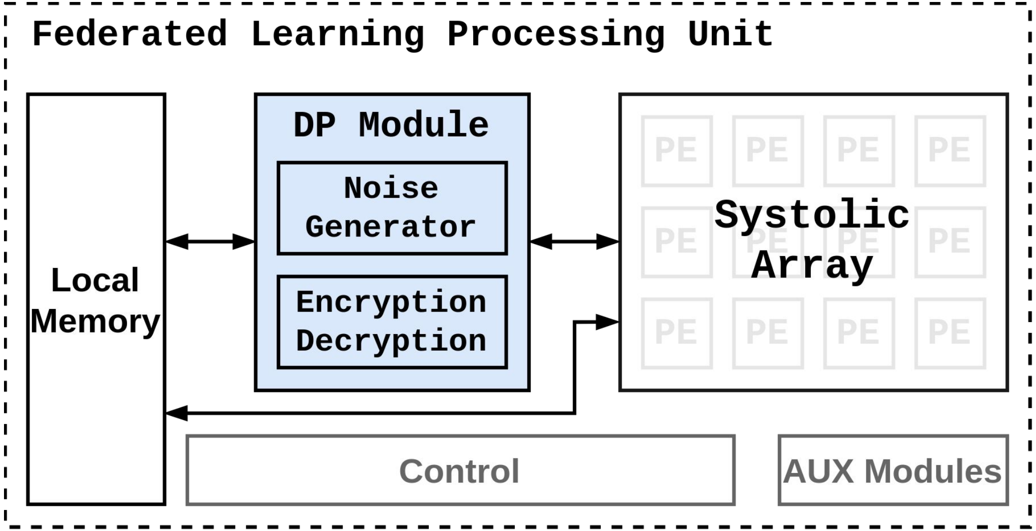


Federated Learning Processing Unit

Possible design choices for **Differential Privacy (DP) module**:

- Generation of noise within the chip (Noise Generator)
- Encryption/Decryption acceleration

- DP Module can be used in several ways:
  - Add noise to input/output data
  - Add noise to model weights



Conclusion

Conclusion

- Federated Learning is a promising approach for **distributed training** of machine learning models
- One of the most popular technique to ensure user data privacy is **differential privacy**
- One of the key challenges is to accelerate training
- **Federated Learning Processing Unit** (FLPU) can be implemented to speed up training process under differential privacy conditions
- FLPU is composed of **several modules**, each of which requires detailed analysis to be designed

Accelerating Differential Privacy-Based Federated Learning Systems

Mirco Mannino, Alessio Medaglini, Biagio Peccerillo, and Sandro Bartolini

Thank You!
Any Questions or Suggestions?

UNIVERSITÀ DI SIENA 1240

Department of Information Engineering and Mathematics
University of Siena, Italy

e-mail: mannino@diism.unisi.it

IARIA

18th International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP) 2024
Special Session: Hardware Accelerators and Accelerated Programming (HAAP)

29th September, 2024