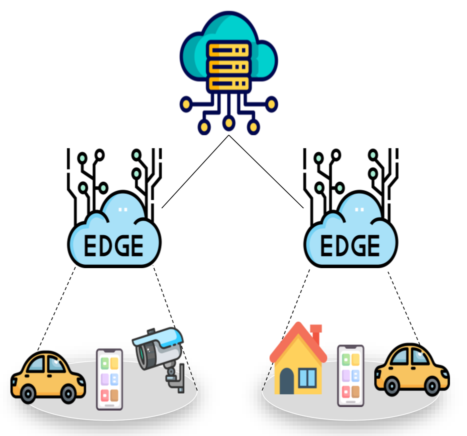
ESFL: Edge-assisted Split Federated Learning

Van An Le, Jason Haga, Yusuke Tanimura, Truong Thao Nguyen
The National Institute of Advanced Industrial Science and Technology (AIST), Japan

Federated Learning (FL) has played a critical role in supporting the development of AI-based privacy-sensitive applications. We introduce ESFL, a novel FL scheme addressing the challenges of developing FL in the Thing-Edge-Cloud environment.

Challenges (What?)



- Communication bottleneck at cloud server due to the large number of devices.
- Resource constraints at IoT devices.
- Low accuracy due to heterogeneous data distributions (non-IID data).

Research Approach (How?)

Centralized training of a high-capacity model on the cloud.

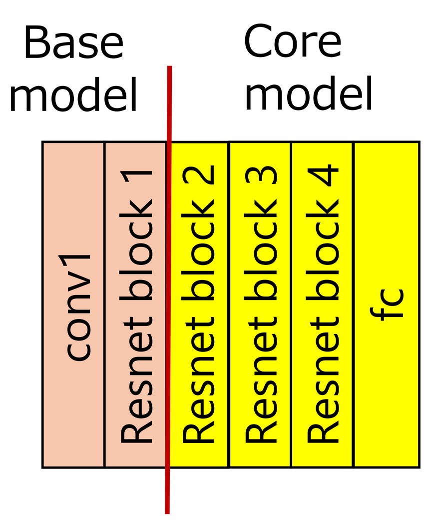
- Collecting data from multiple IoT devices to cloud servers to mitigate the non-IID issues.
  - Privacy → **collecting feature vector of the data.**
- Perform the feature vector preprocessing at edge servers to reduce communication load to the cloud.

ESFL Training Framework

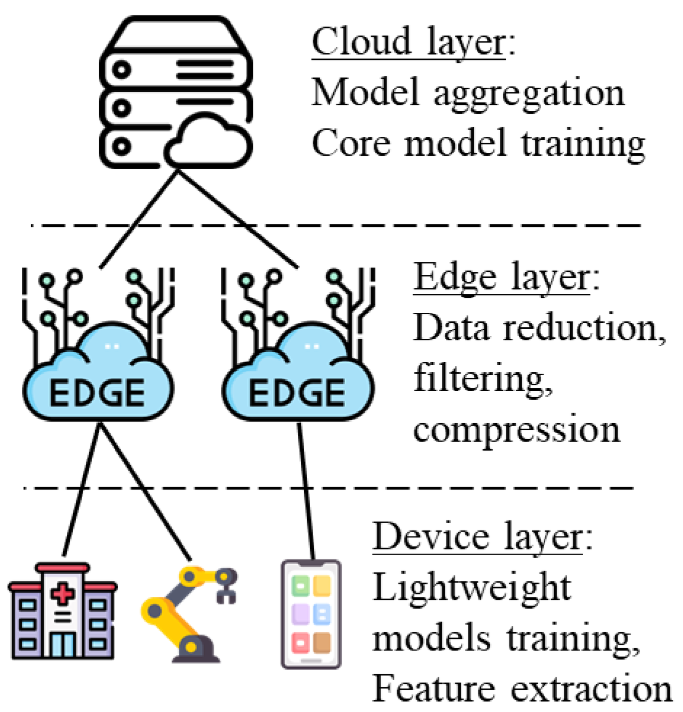
A. Overall System

Original model is divided into a lightweight Base model & Core model.

- Base and Core models are trained separately, avoiding traffic communication congestion.

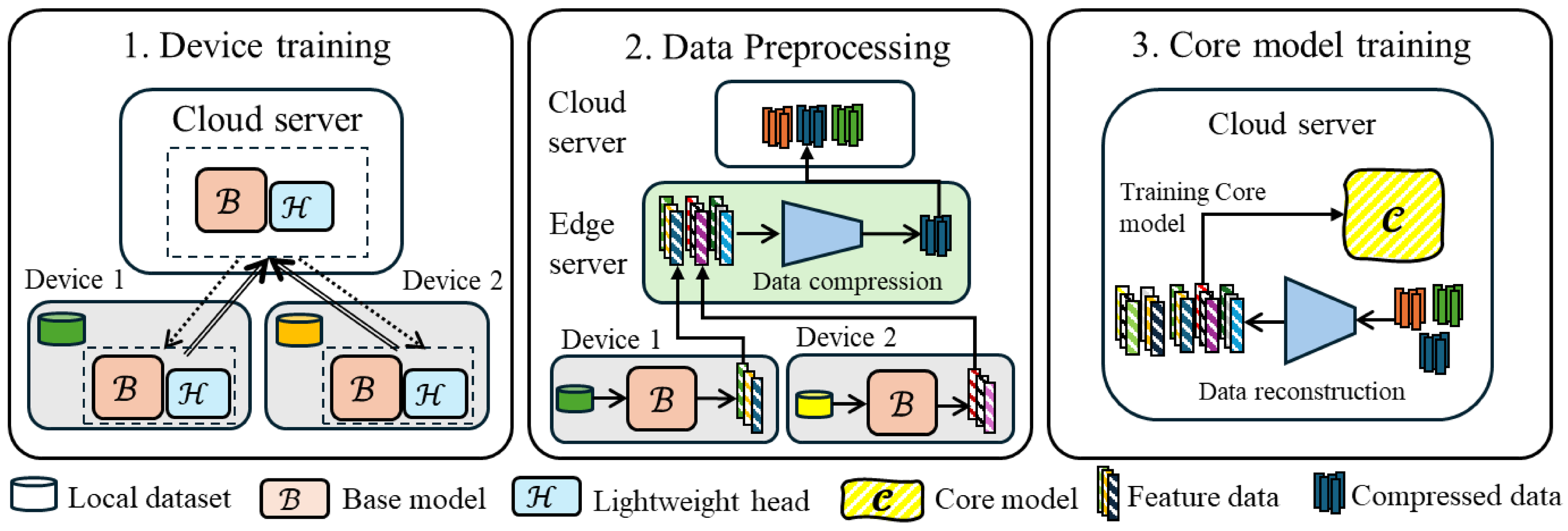


An example of a split model using ResNet-18



B. Training process

- Base model is trained on devices with a lightweight head model using FedAvg.
- Base model is then utilized to extract feature representations from local data, which are then sent to the cloud for training Core model.
- Performs the edge-side pre-processing of feature data before transmitting to the cloud to reduce the communication overhead by **(1) Random select p% of data and (2) data compression**, e.g., using ZeroQuantV2.



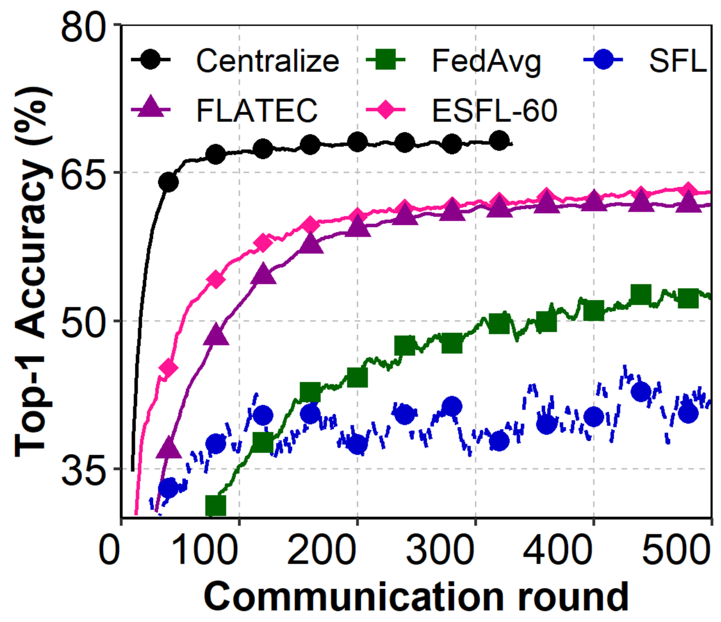
Preliminary Result

A. Experimental settings

Dataset	Cifar100
Non-IID	Dirichlet (0.05)
No. devices	100
No. edges	20
Training	500 rounds
Model	ResNet-18

Baseline approaches:
- **FedAvg**: training full model at devices.
- **SFL**: partially training at cloud.
- **FLATEC**: training at edge servers

B. Higher testing accuracy & less communication to the cloud



Method	Acc. (%)	Traffic
FedAvg	56.55	898
SFL	51.21	13110
FLATEC	62.19	334
ESFL-20	59.08	245
ESFL-40	63.07	449
ESFL-60	63.66	654
ESFL-80	63.21	858
ESFL-100	63.43	1063
ESFL-NQ	63.92	2658

- **ESFL-p**: ESFL with p% features data are randomly selected at edge servers.
- **ESFL-NQ**: ESFL without quantization.
- **Traffic**: average traffic load to the cloud per round in MB.

- ESFL-60 achieves the best trade-off, reaching a 63.66% accuracy.
- ESFL exhibits faster convergence than competing baselines.
- ESFL robust across different sampling ratios p
- ZeroQuantV2 provides substantial compression with minimal performance degradation.

Future Work

- Study the impact of ESFL on different non-IID scenarios.
- Study the impact of data selection mechanisms at edge-servers on the performance of ESFL.
- Study the computational overhead, e.g., introduced by data compression/decompression.

References

[1] **SFL:** C. Thapa et al., "Splitfed: When federated learning meets split learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, 2022, pp. 8485–8493.
[2] **FLATEC**: Van An Le et al., "Flatec: An efficient federated learning scheme across the thing-edge-cloud environment," Future Generation Computer Systems, vol. 175, p. 108073, 2026.
[3] **Zeroquant-v2**: Z. Yao et al., "Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation," arXiv preprint arXiv:2303.08302, 2023.
[4] **FedAvg**: B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.

This paper is based on results obtained from the project, "Research and Development Project of the Enhanced infrastructures for Post-5G Information and Communication Systems" (JPNP20017), commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Van An LE
Intelligent Platforms Research Institute
Contact: an.le@aist.go.jp

Create the Future, Collaborate Together