Semi-Decentralized Federated Edge Learning for Fast Convergence on Non-IID Data

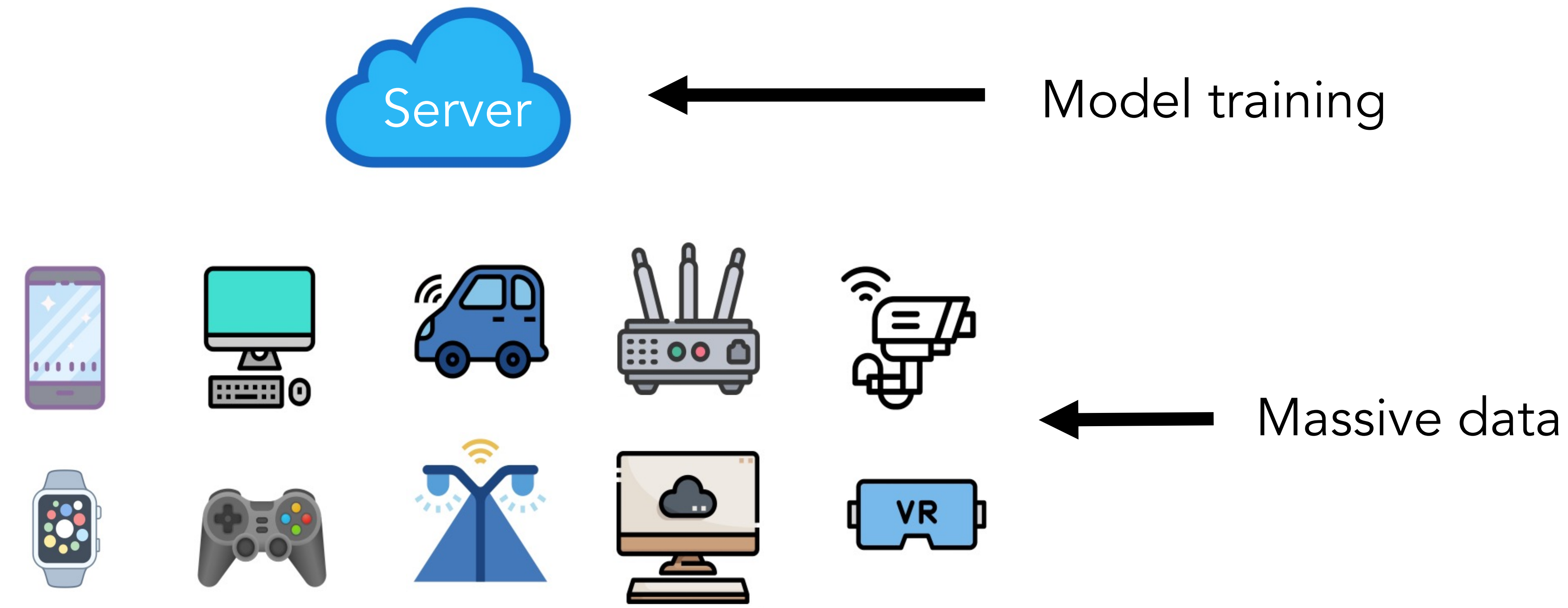
Authors: *Yuchang Sun*, Jiawei Shao, Yuyi Mao, Jessie Hui Wang, and Jun Zhang

WCNC 2022

Content

- Background & Motivation
- Existing works
- Approach
- Results
- Conclusions

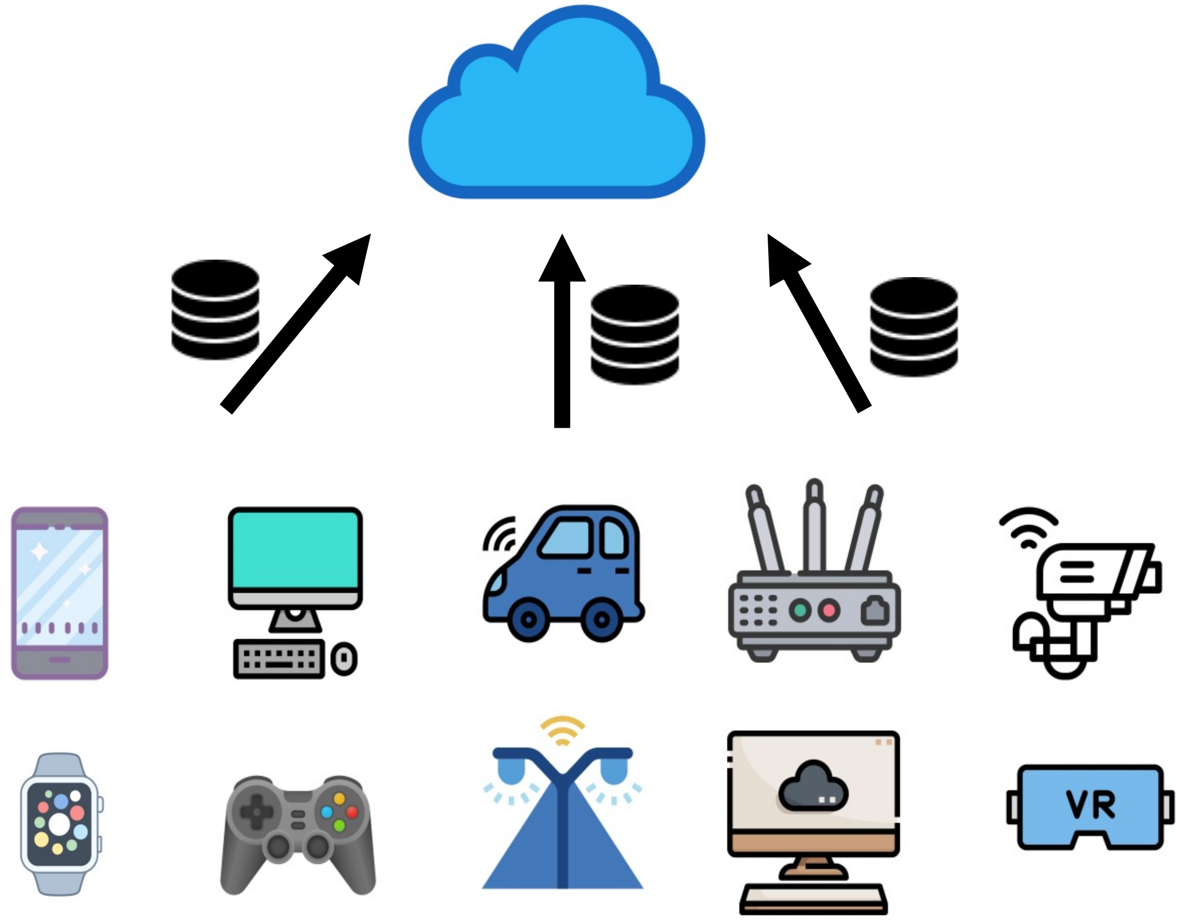
Background



More than 30 billion IoT devices by 2025 [1].

[1] K. L. Lueth, "State of the IoT 2020: 12 billion IoT connec- tions, surpassing non-IoT for the first time." Nov. 2020. [Online]. Available: https://iot-analytics.com/state-of-the-iot-2020-12-billion-iot- connections- surpassing- non- iot- for- the- first- time/

Background

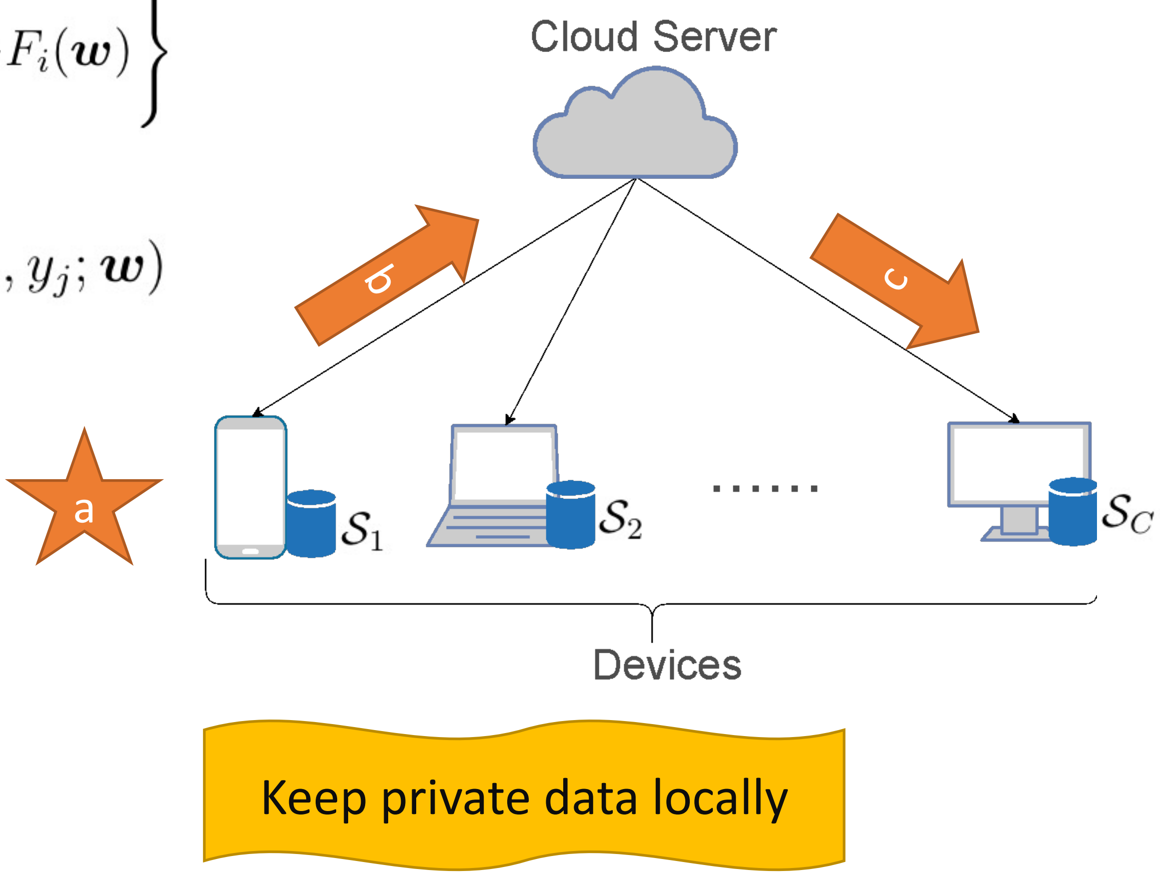


Upload data?

Privacy leakage

Background: Federated Learning [2]

- Global objective $\min_{\boldsymbol{w}\in\mathbb{R}^M} \left\{ F(\boldsymbol{w}) \triangleq \sum_{i\in\mathcal{C}} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} F_i(\boldsymbol{w}) \right\}$
- Local objective $F_i(\boldsymbol{w}) \triangleq \frac{1}{|\mathcal{S}_i|} \sum_{j\in\mathcal{S}_i} f(\boldsymbol{x}_j, y_j; \boldsymbol{w})$
- In each training round:
  a. local update
  b. model aggregation
  c. broadcast



[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Ft. Lauderdale, FL, USA, Apr. 2017.

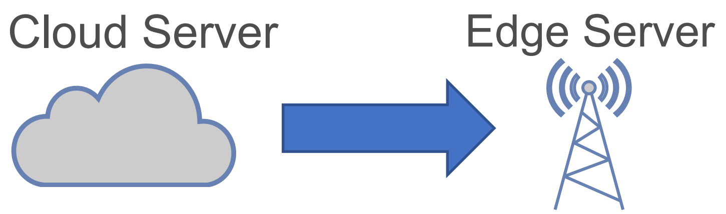
Motivation: Improve Training Efficiency

- FL task comprises a massive number of devices.
  - Local training requires great computation resources.
  - Slow devices may prolong the training time.
- Communication between devices and the Cloud takes a long time!
  - Some devices may have unfavorable channel conditions.

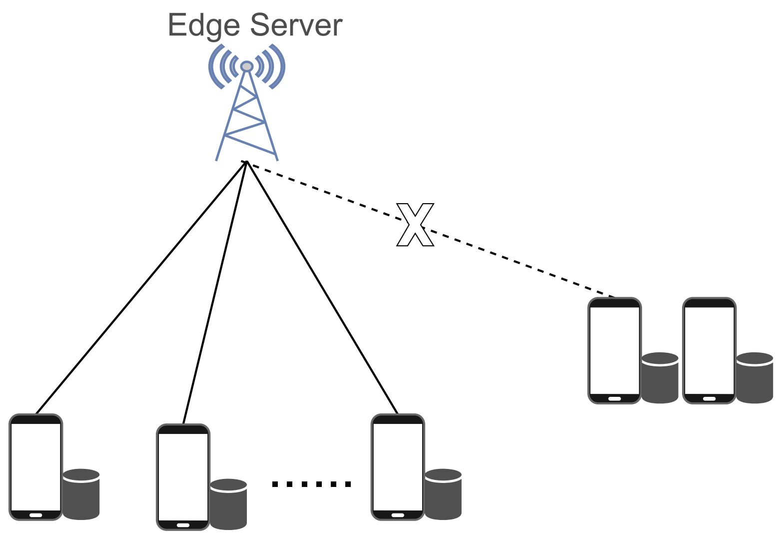
How to reduce training time?

Existing Works: FEEL

- Federated Edge Learning (FEEL) [3]
  - Push the aggregation task to the *edge*.

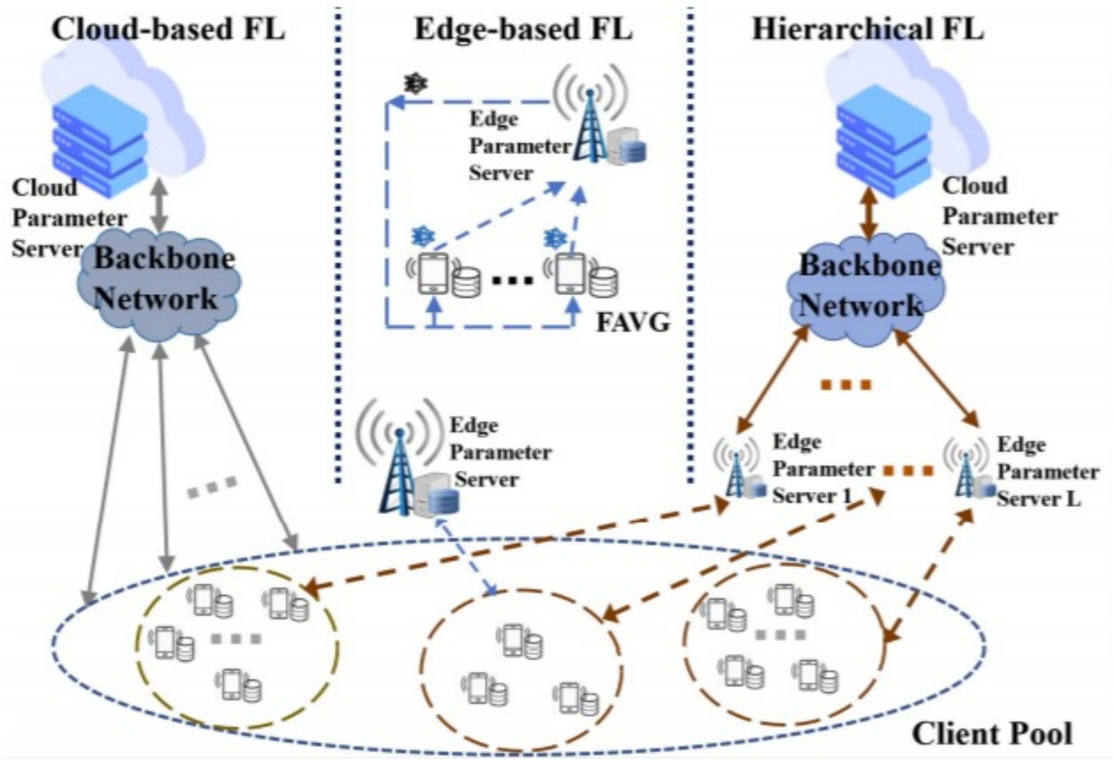


- New challenges
  - Limited coverage of one single edge server.
  - Less training data than Cloud-based FL.



[3] W. Y. B. Lim et al., “Federated learning in mobile edge networks: A comprehensive survey,” IEEE Commun. Surveys Tuts., vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.

Existing Works: Hierarchical FL [4]

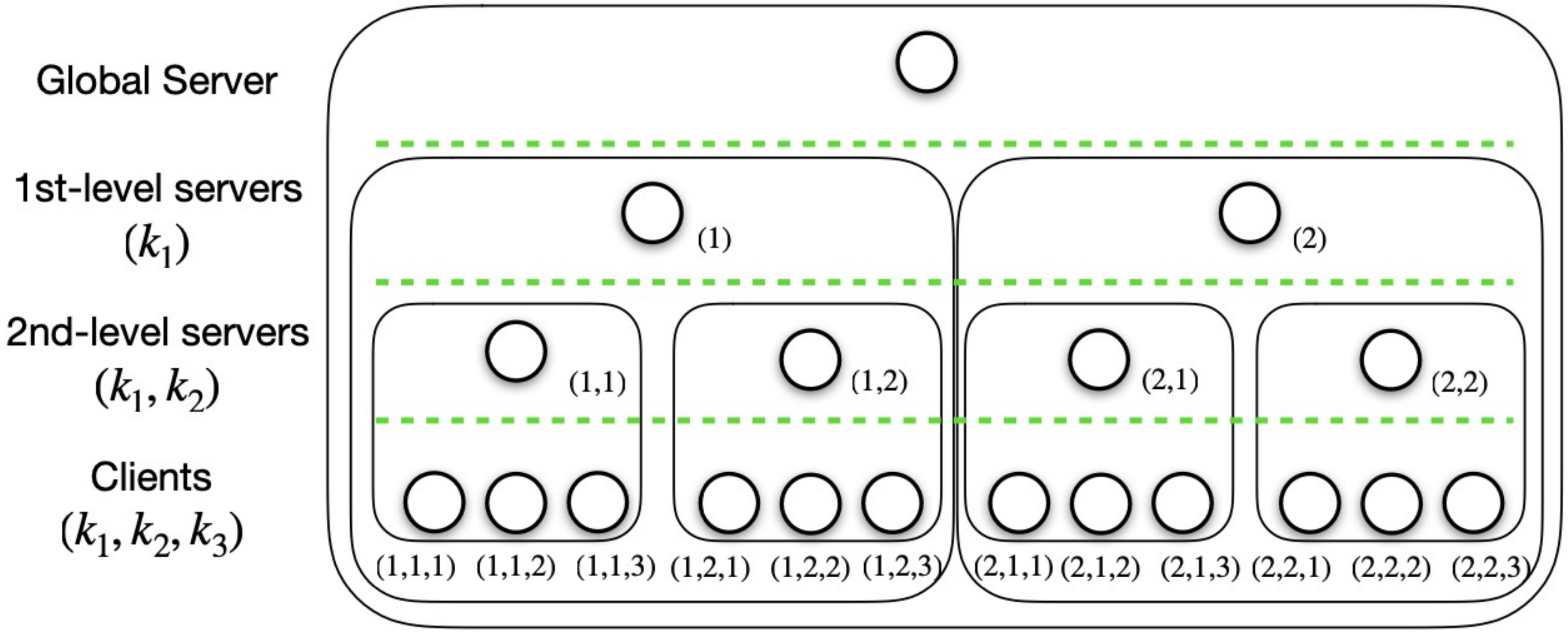


- Utilize multiple edge servers to accelerate model training.
- Communication latency with the Cloud is still high!

[4] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in Proc. IEEE Int. Conf. Commun.(ICC), Dublin, Ireland, Jun. 2020.

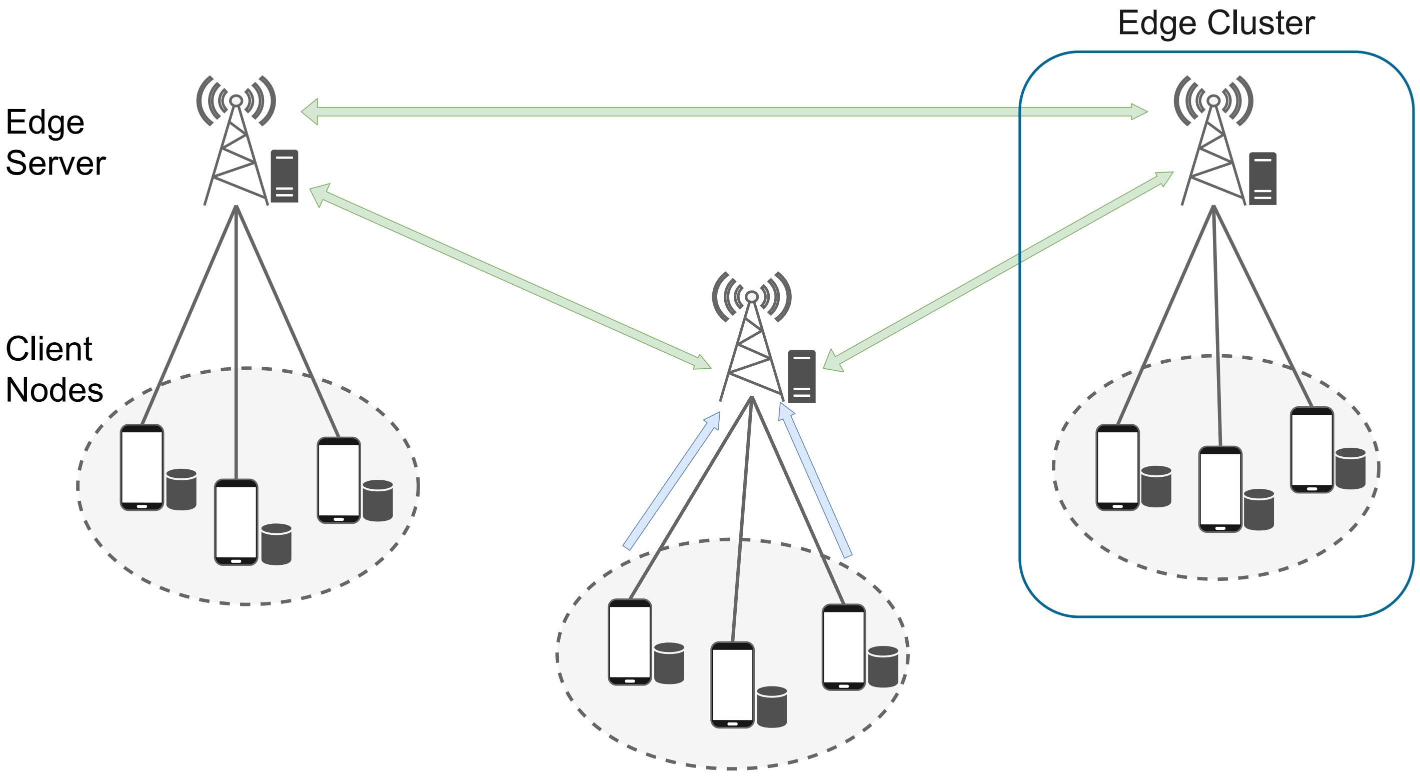
Existing Works: Hierarchical federated SGD [5]

- Extend [4] to a multi-level case.



[5] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Local averaging helps: Hierarchical federated learning and convergence analysis." [Online]. Available: https://arxiv.org/pdf/2010.12998.pdf

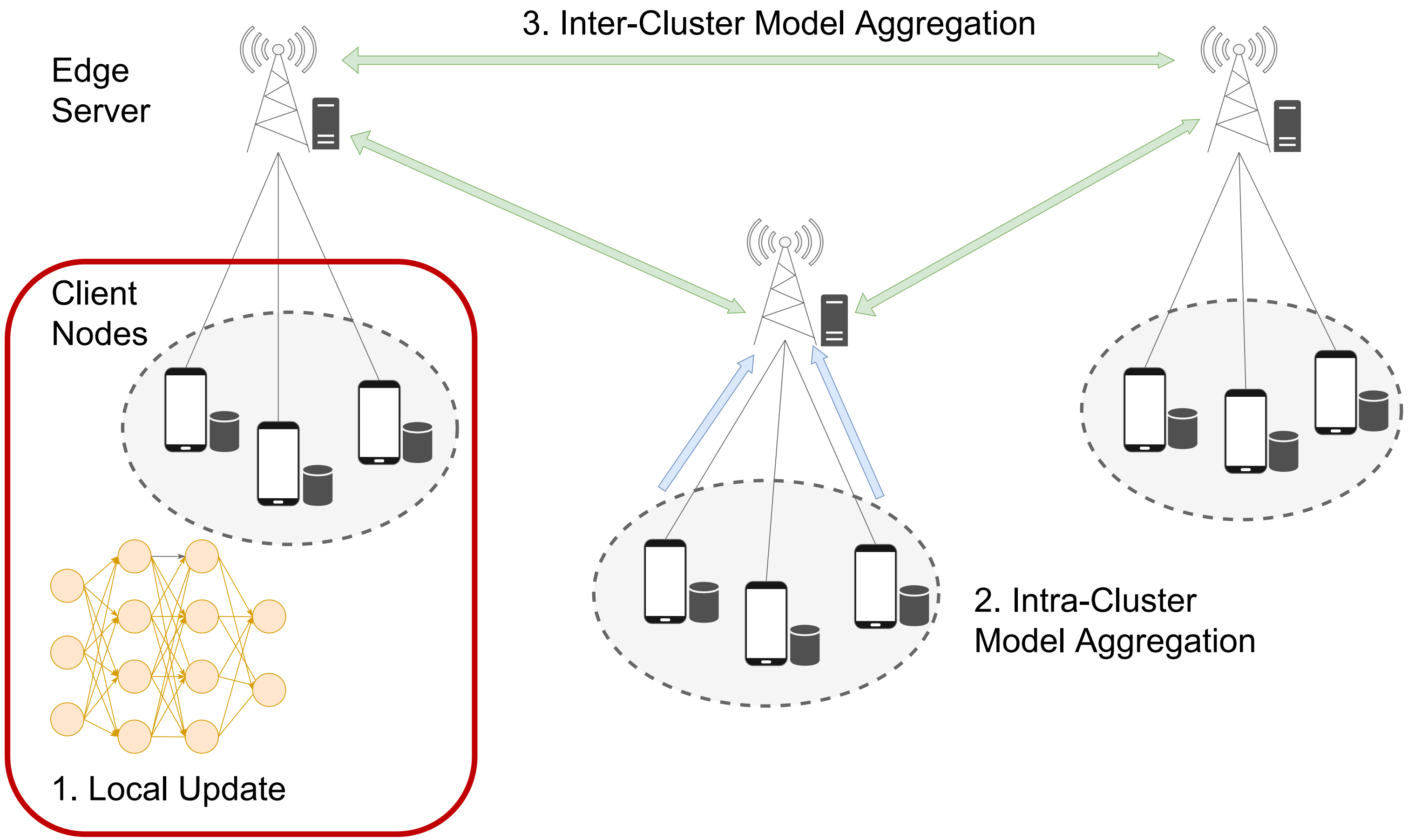
Approach: SD-FEEL



- Efficient communication among edge servers.
- Servers collaborate with each other to get more information.
- No additional computation on clients.

Note: SD-FEEL is the abbreviation for Semi-decentralized federated edge learning.

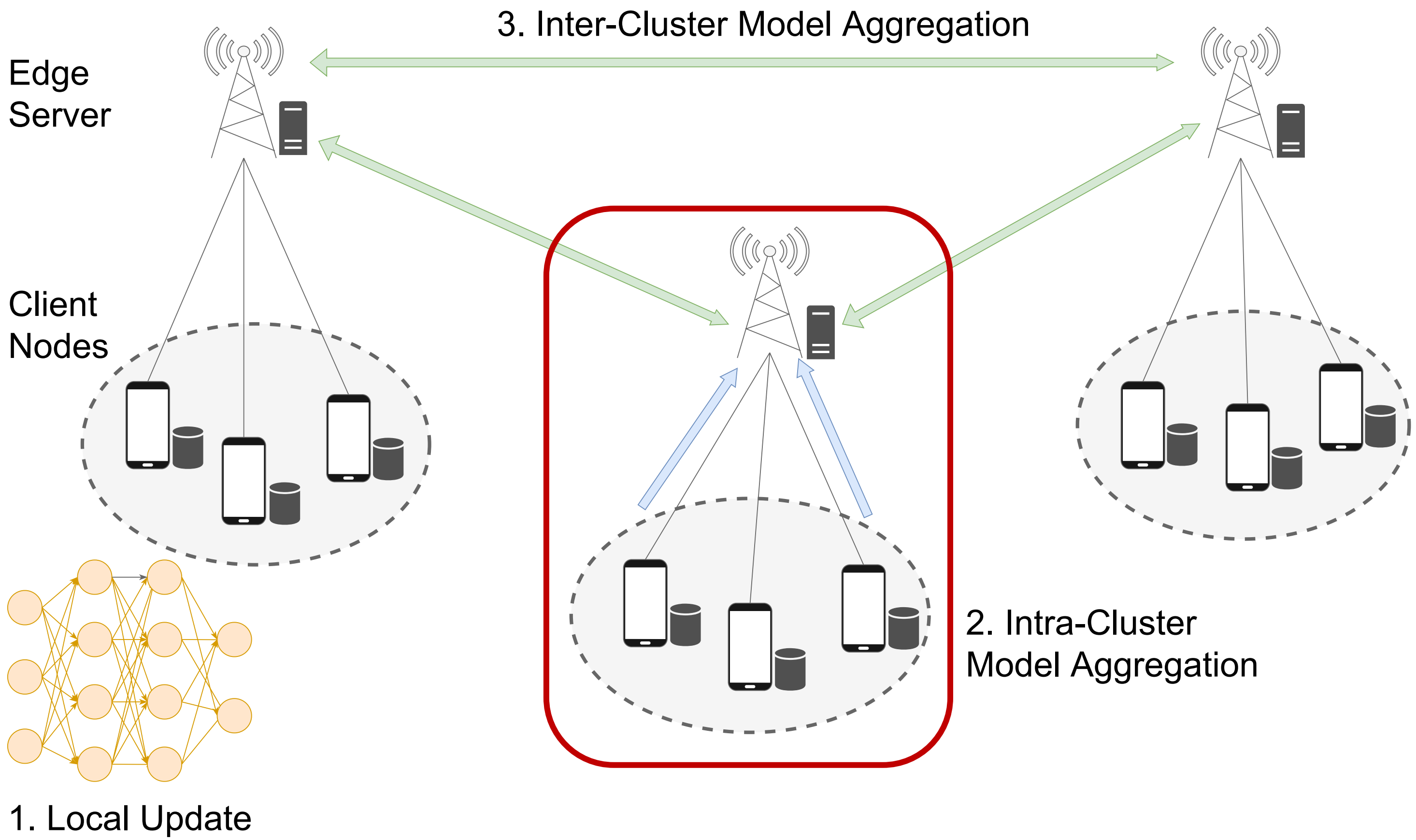
Approach: Training Process



(1) Local Updates

$$\boldsymbol{w}_k^{(i)} \leftarrow \boldsymbol{w}_{k-1}^{(i)} - \eta g(\boldsymbol{\xi}_k^{(i)}; \boldsymbol{w}_{k-1}^{(i)}), i \in \mathcal{C}$$

Approach: Training Process

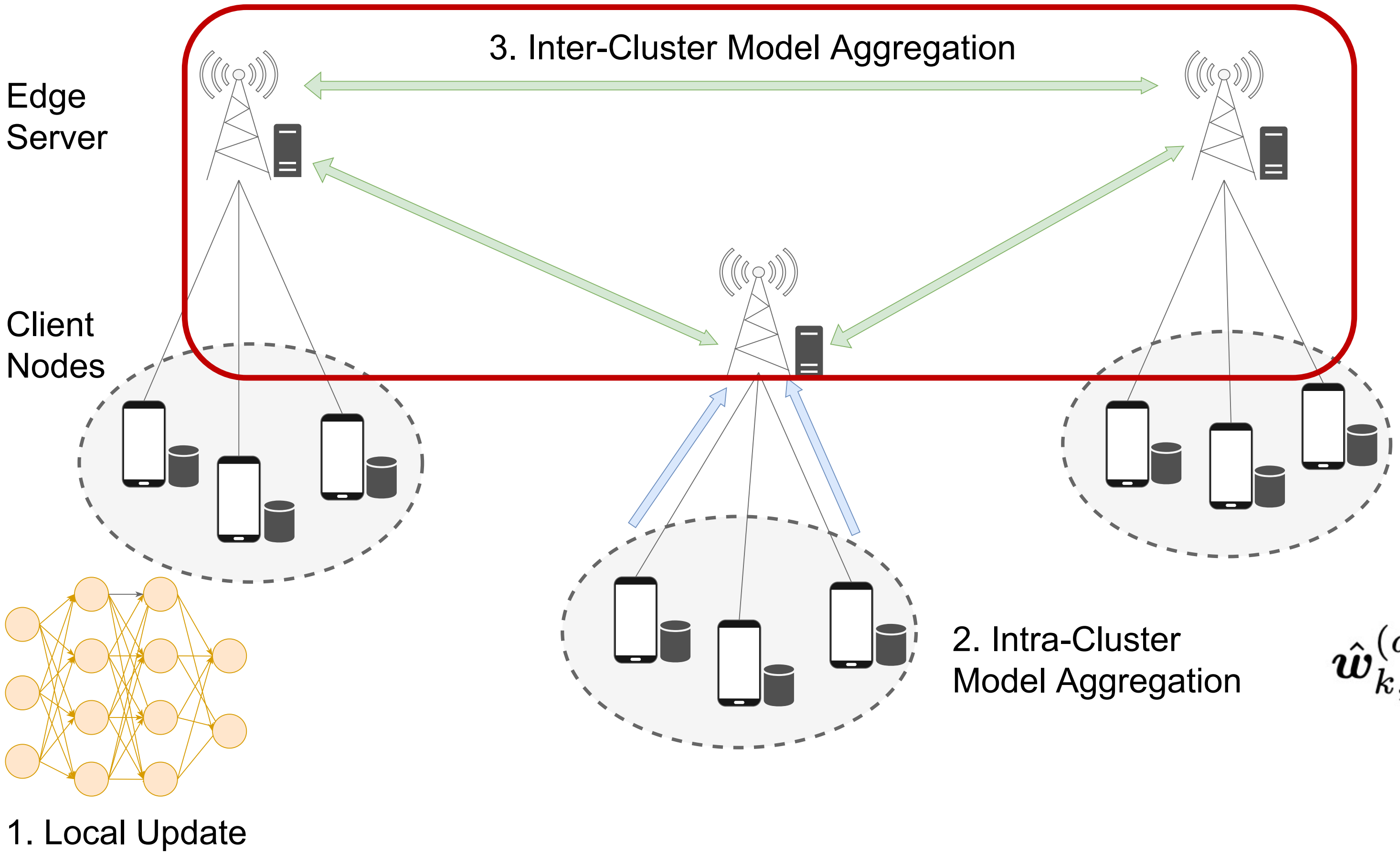


(2) *Intra-Cluster* Model Aggregation

- Scheduled after every τ_1 local epochs
- Weighted average

$$\tilde{\boldsymbol{w}}_k^{(d)} \leftarrow \sum_{i \in \mathcal{C}_d} \frac{|\mathcal{S}_i|}{|\tilde{\mathcal{S}}_d|} \boldsymbol{w}_k^{(i)}, d \in \mathcal{D}$$

Approach: Training Process

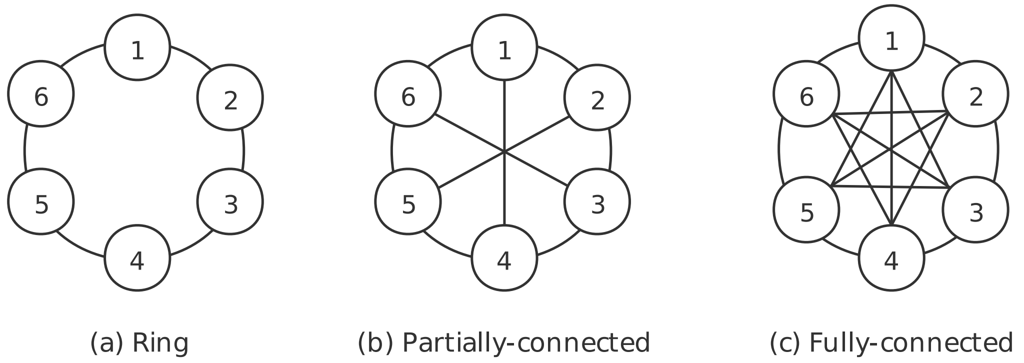


(3) *Inter-Cluster* Model Aggregation

- Scheduled after every $\tau_1 \tau_2$ local epochs
- Perform α times of model exchanges

$$\hat{\boldsymbol{w}}_{k,l}^{(d)} \leftarrow \sum_{j \in \mathcal{N}_d \bigcup \{d\}} p_{j,d} \hat{\boldsymbol{w}}_{k,l-1}^{(j)}, l = 1, 2, \ldots, \alpha$$

Approach



- Multi-level SGD [6] investigated a similar architecture.

- It assumed only one round of communication among edge servers.
  - May cause may model inconsistency and degrade model performance.

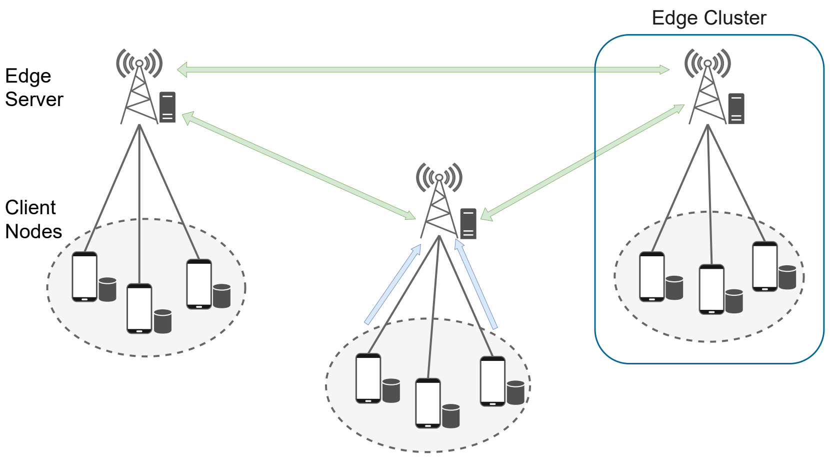
- Convergence analysis is limited to IID* local training data.

*Independent and identically distributed.

[6] T. Castiglia, A. Das, and S. Patterson, "Multi-level local SGD: Dis- tributed SGD for heterogeneous hierarchical networks," in Proc. Int. Conf. Learn. Repr. (ICLR), Virtual Event, May 2021.

Results: Theoretical Challenge

- Expected loss change involves:
  - Two levels of model aggregations
  - Decentralized topology among edge servers
  - Multiple rounds of inter-server communication



- The effect of non-IID data
  - Mismatch between local objective and global objective.

$$\nabla f_i(\boldsymbol{w}) \neq \nabla f(\boldsymbol{w})$$

[6] T. Castiglia, A. Das, and S. Patterson, "Multi-level local SGD: Dis- tributed SGD for heterogeneous hierarchical networks," in Proc. Int. Conf. Learn. Repr. (ICLR), Virtual Event, May 2021.

Results: Theoretical Convergence

- Model evolution

Lemma 1. *The local models evolve according to the following expression:*

$$\mathbf{W}_{k+1} = (\mathbf{W}_k - \eta \mathbf{G}_k)\mathbf{T}_k,\ k = 1, 2, \ldots, K, \tag{10}$$

where

$$\mathbf{T}_k = \begin{cases} \mathbf{VB}, & \textit{if} \bmod(k,\tau_1) = 0 \textit{ and} \bmod(k,\tau_1\tau_2) \neq 0, \\ \mathbf{VP}^{\alpha}\mathbf{B}, & \textit{if} \bmod(k,\tau_1\tau_2) = 0, \\ \mathbf{I}, & \textit{otherwise.} \end{cases} \tag{11}$$

Results: Theoretical Convergence

- Define a model $\boldsymbol{u}_k \triangleq \sum_{i \in \mathcal{C}} m_i \boldsymbol{w}_k^{(i)}$

$$m_i \triangleq \frac{|\mathcal{S}_i|}{|\mathcal{S}|}$$

- The expected change in consecutive iterations:

$$\mathbb{E}[F(\boldsymbol{u}_{k+1})] - \mathbb{E}[F(\boldsymbol{u}_k)] \leq -\frac{\eta}{2}\mathbb{E}\left[\|\nabla F(\boldsymbol{u}_k)\|_2^2\right] + \frac{\eta^2 L}{2}\sum_{i \in \mathcal{C}} m_i^2 \sigma^2 - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2}\right) J_k + \boxed{\frac{\eta L^2}{2}\mathbb{E}\left[\|\mathbf{W}_k(\mathbf{I}-\mathbf{M})\|_{\mathbf{M}}^2\right]}, \quad (12)$$

Results: Theoretical Convergence

- The deviation of the local models from their mean:

Lemma 2. *With Assumption 1, we have:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\mathbf{W}_k(\mathbf{I}-\mathbf{M})\|_{\mathbf{M}}^2\right] \leq \frac{8\eta^2 V_2}{K}\sum_{k=1}^{K} J_k + 2\eta^2 V_1\sigma^2 + 8\eta^2 V_2\kappa^2, \tag{13}$$

where $\zeta = |\lambda_2(\mathbf{P})| \in [0,1)$, $\Lambda \triangleq \frac{\zeta^{2\alpha}}{1-\zeta^{2\alpha}} + \frac{2\zeta^{\alpha}}{1-\zeta^{\alpha}} + \frac{\zeta^{2\alpha}}{(1-\zeta^{\alpha})^2}$, $V_3 \triangleq \tau_1\tau_2\left(\tau_1\tau_2\Lambda + \frac{\tau_1\tau_2-1}{2}\frac{2-\zeta^{\alpha}}{1-\zeta^{\alpha}}\right)$, $V_1 \triangleq \left(\tau_1\tau_2\frac{\zeta^{2\alpha}}{1-\zeta^{2\alpha}} + \frac{\tau_1\tau_2-1}{2}\right)/(1-16\eta^2L^2V_3)$, *and* $V_2 \triangleq V_3/(1-16\eta^2L^2V_3)$.

Results: Theoretical Convergence

Theorem 1. *If the learning rate η satisfies:*

$$1 - \eta L - 8\eta^2 L^2 V_2 \geq 0, 1 - 16\eta^2 L^2 V_3 > 0, \tag{14}$$

we have:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\boldsymbol{u}_k)\|_2^2\right] \leq \boxed{\frac{2\Delta}{\eta K} + \eta L \sum_{i \in \mathcal{C}} m_i^2 \sigma^2} + \boxed{2\eta^2 L^2 V_1 \sigma^2 + 8\eta^2 L^2 V_2 \kappa^2}, \tag{15}$$

where $\Delta \triangleq \mathbb{E}[F(\boldsymbol{u}_1)] - \mathbb{E}[F(\boldsymbol{u}^*)]$ *and* $\boldsymbol{u}^* \triangleq \arg\min_{\boldsymbol{w}} F(\boldsymbol{w})$.

Existed in centralized SGD

Additional error

- Detailed proof [7]

[7] Y. Sun, J. Shao, Y. Mao, J. H. Wang, and J. Zhang, "Semi-decentralized federated edge learning for fast convergence on non-IID data." [Online]. Available: https://arxiv.org/pdf/2104.12678.pdf

Results: Insights from Convergence

1. When $\tau_1 = \tau_2 = 1$ and $\zeta^{\alpha} = 0$, the convergence result in Theorem 1 reduces to that of the fully synchronous SGD algorithm [8].

2. More frequent intra-/inter-cluster model aggregation

3. For inter-server communication:
    - a more connected topology
    - increasing the communication overhead

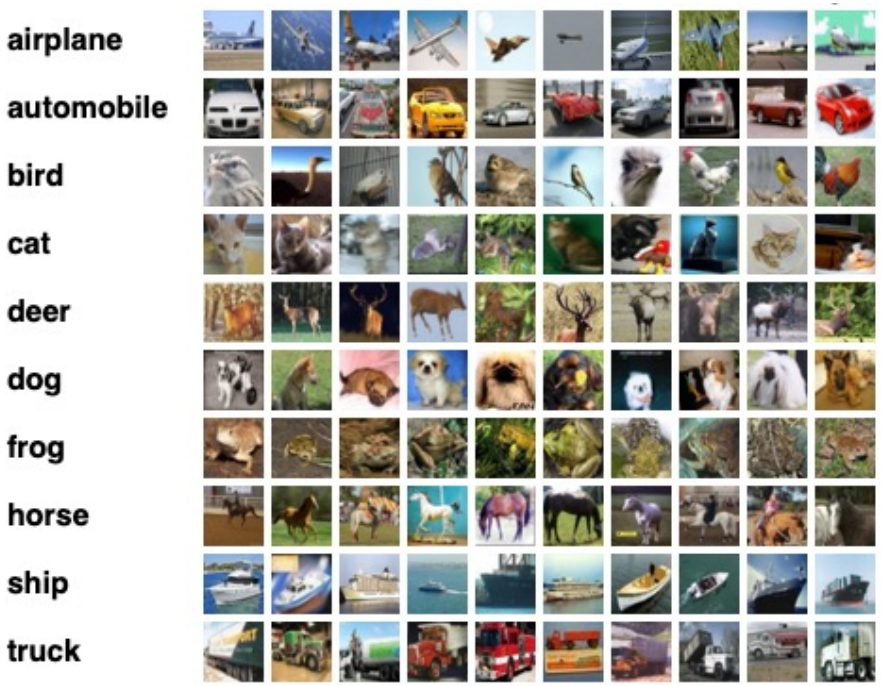
Faster convergence

[8] L.Bottou,F.E.Curtis,andJ.Nocedal,"Optimizationmethodsforlarge- scale machine learning," SIAM Rev., vol. 60, no. 2, pp. 223-311, Aug. 2018.

Results: Experimental Setup

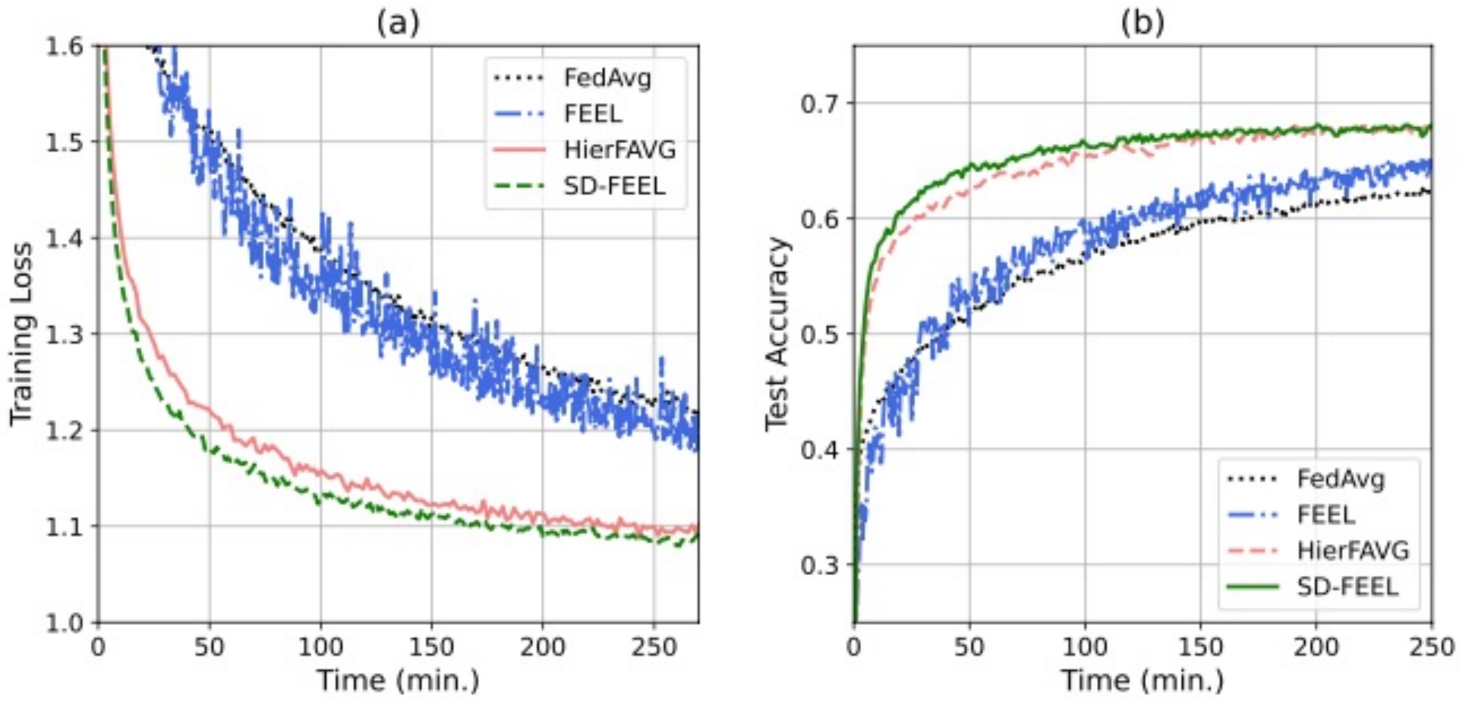
- 50 clients, 10 edge servers.
- CIFAR-10 dataset + CNN model [4]
- Data partition: Dirichlet distribution [9]

- Baselines:
  - FedAvg [2]
  - FEEL with partial participation [3]
  - HierFAVG [4]



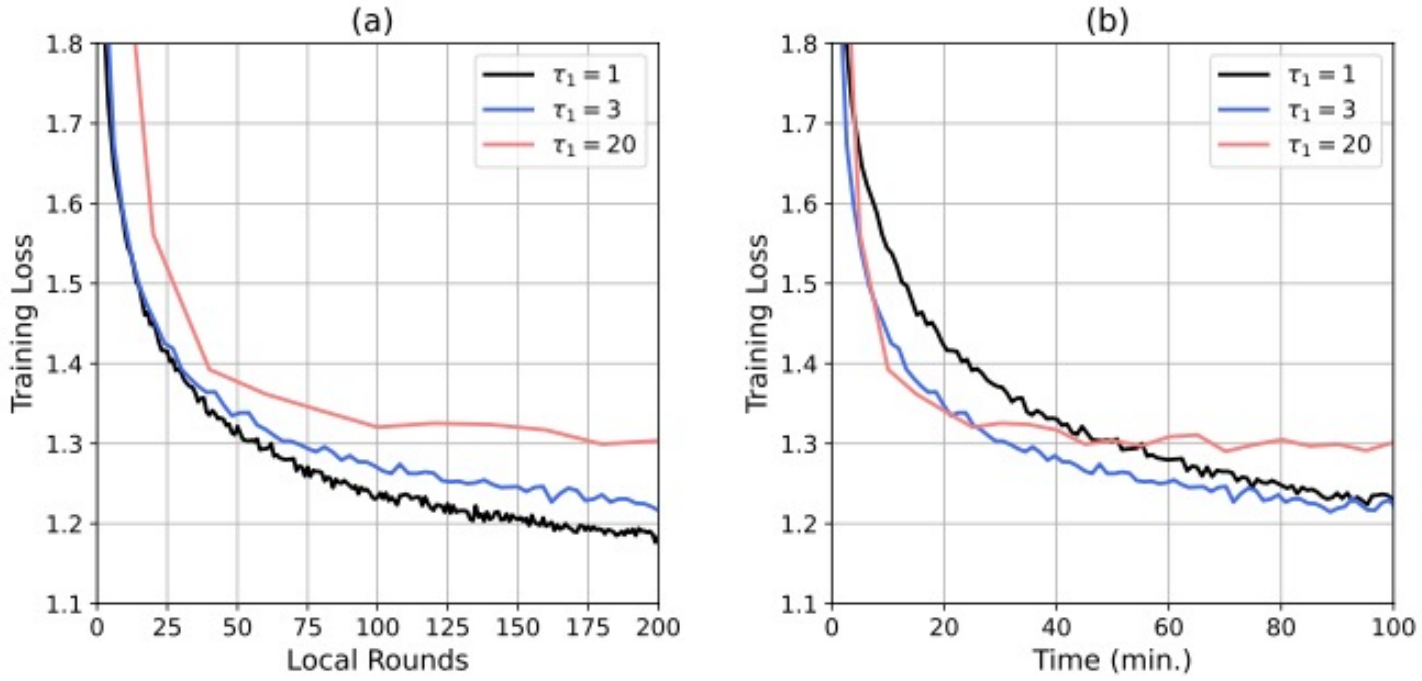
[9] H. Wang, M. Yurochkin, Y. Sun, D. S. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in Proc. Int. Conf. Learn. Repr. (ICLR), Addis Ababa, Ethiopia, Apr. 2020.

Results: Convergence Performance



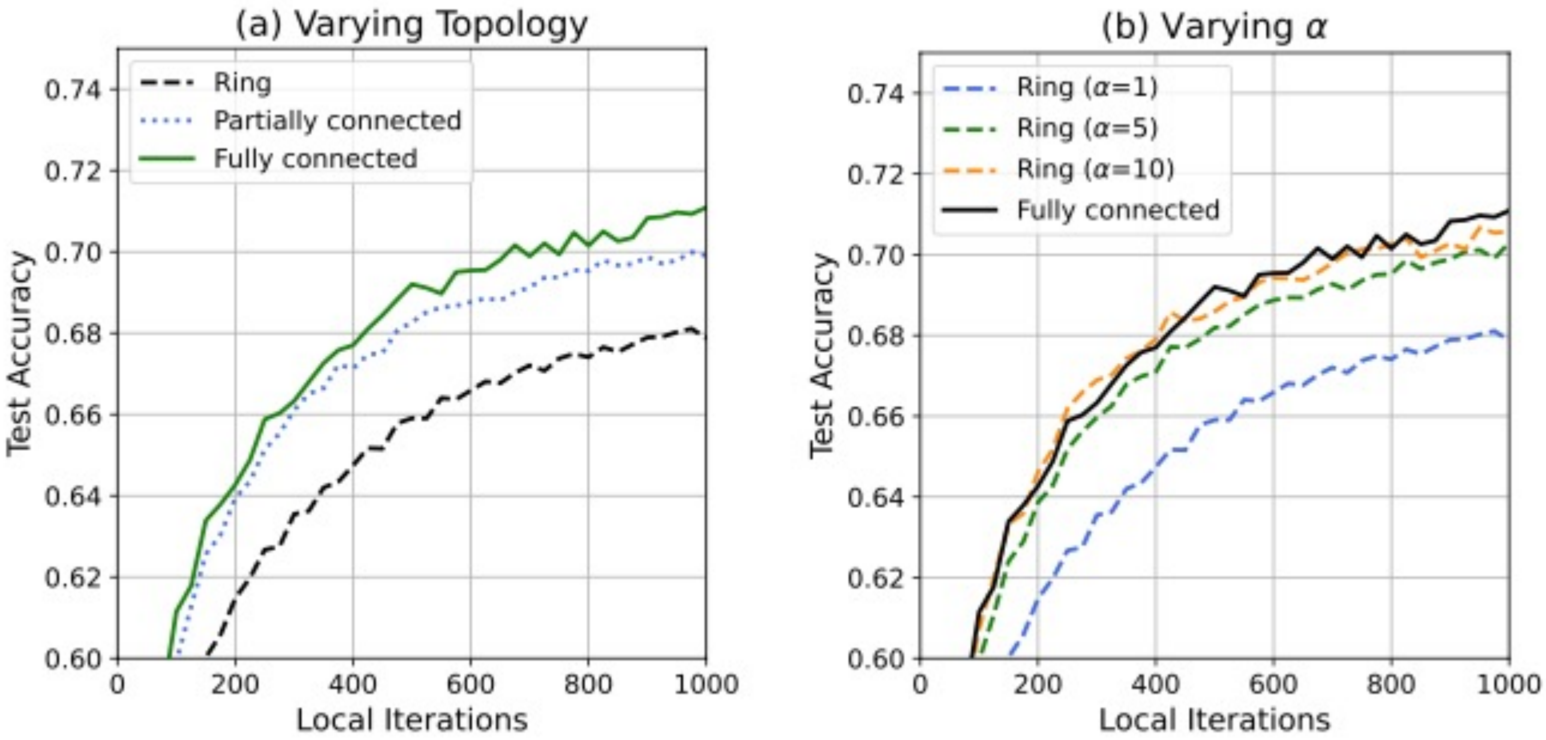
- Within the given training time, SD-FEEL converges fast and has a higher accuracy.
- Communication among edge servers is more efficient.

Results: Ablation Study of τ_1 and τ_2



- Considering training rounds, more frequent aggregation is preferred.
- Within the same training time, $\tau_1 = 3$ achieves the minimum training loss.

Results: Ablation Study of Topology



- A more connected network topology achieves a higher test accuracy.
- More information is collected from neighboring edge clusters.

Conclusions

- Investigated *semi-decentralized federated edge learning* (SD-FEEL).
  - Proved convergence analysis (on non-IID data)
  - Empirically demonstrated the high training efficiency.
- Provided guidelines on selection of system parameters.
  - Larger aggregation frequency improves convergence speed but incurs communication overhead.
  - Multiple times of inter-server communication speeds up convergence.
- Future works: consider scenarios with device heterogeneity.

Thank you!