IJCNN, International Joint Conference on Neural Networks,
WCCI, IEEE World Congress on Computational Intelligence, Padua, Italy 2022

Optimized and Adaptive Federated Learning for Straggler-Resilient Device Selection

Sourasekhar Banerjee, Xuan-Son Vu, and Monowar Bhuyan
{sourasb, sonvx, monowar} @cs.umu.se

Autonomous **D**istributed **S**ystems **L**ab

Department of Computing Science

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

UMEÅ UNIVERSITY

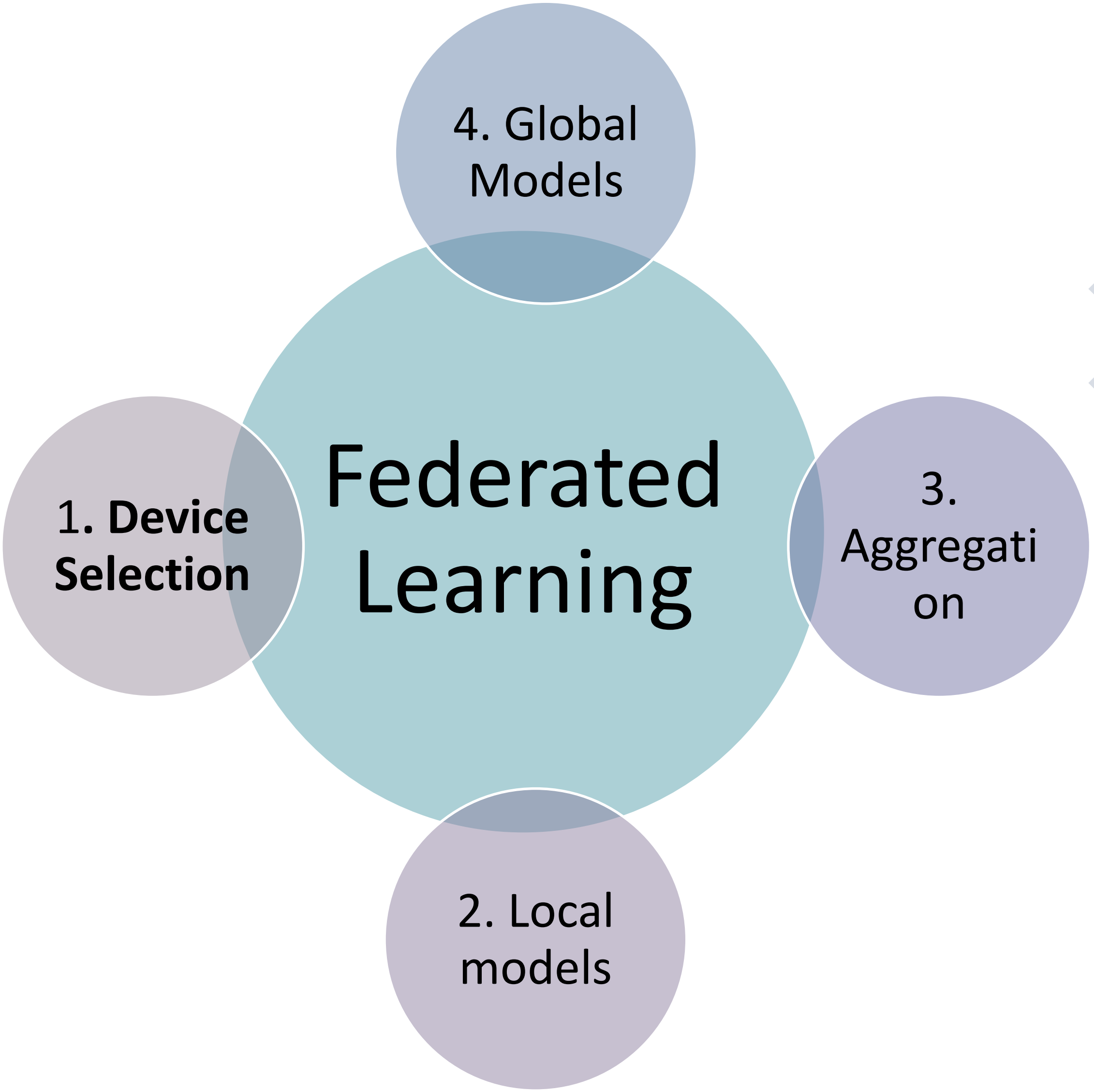
2022-07-13

Outline

- ❖ Introduction
- ❖ Related work
- ❖ Contributions
- ❖ Proposed system model
- ❖ Problem formulations
- ❖ Experiments
- ❖ Results and analysis
- ❖ Conclusions
- ❖ Future scope

Introduction

- ❖ What is federated learning?
  - ❖ Distributed learning.
  - ❖ Collaborative training.
  - ❖ Training on private data.



Introduction

- What is a straggler device ?
  - Low performing device
    - storage
    - communication
    - computation

- Motivations
  - FL is vulnerable to system heterogeneity.
  - When local devices have varying computational, storage, and communication capabilities over time.
  - As a result, the presence of stragglers in the Random selection of devices produces a low convergence rate and high delay in the FL network.

Related Work

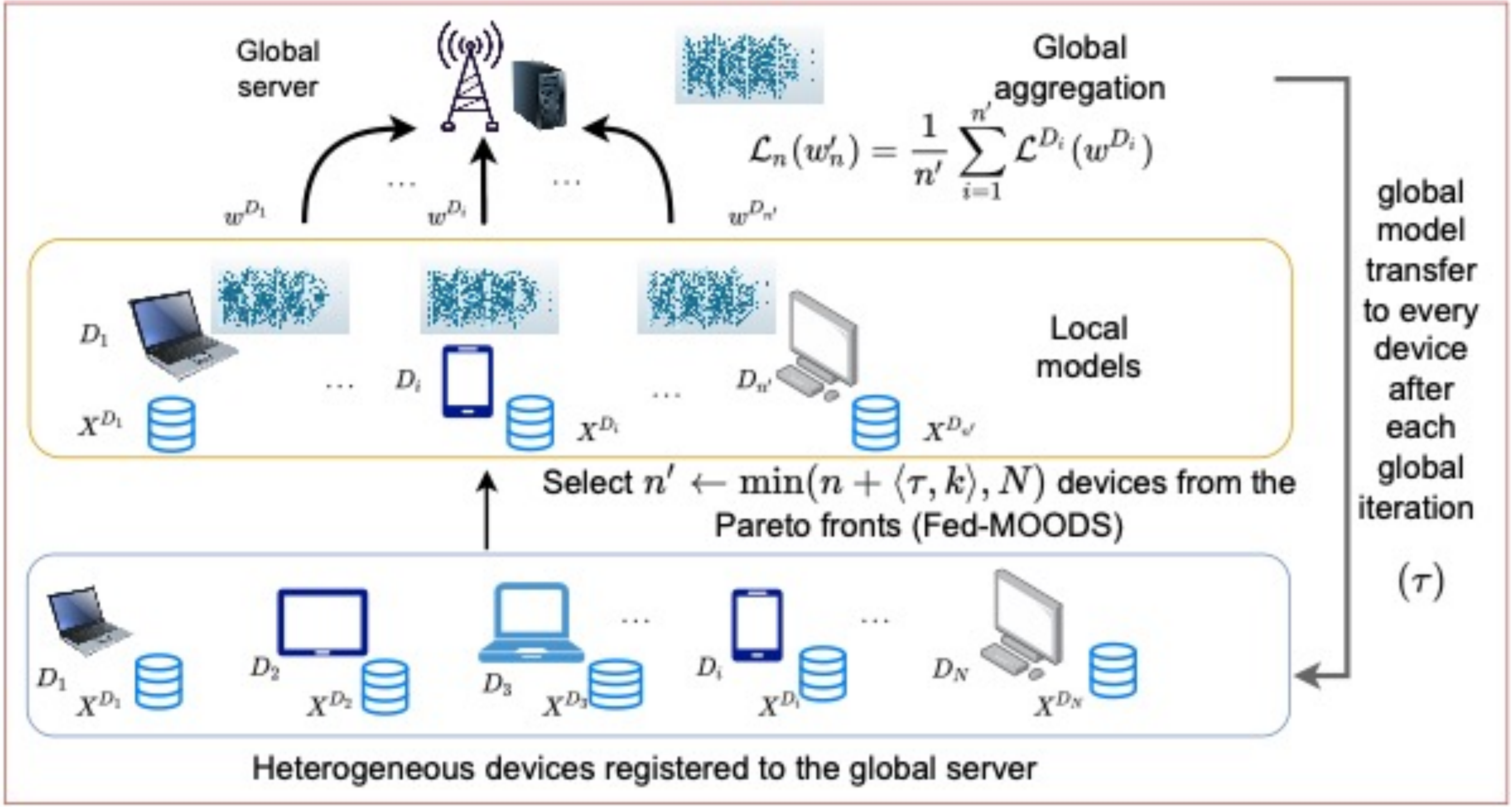
- ❖FLANP, a straggler-resilient adaptive device participation algorithm. (Reisizadeh et al.,2020)
- ❖FedProx, a federated optimization in heterogeneous networks. (Li et al.,2018)
- ❖Accelerated Training via Device Similarity in Federated Learning. (Wang et al.,2021)
- ❖Fair resource allocation in federated learning. (Li et al., 2019)

Contributions

- We introduce **Fed-MOODS**, a straggler-resilient **M**ulti**-O**bjective **O**ptimization-based adaptive prioritized **D**evice **s**election approach to mitigate the system heterogeneity in **Fed**erated learning.

- For each device, Fed-MOODS maximizes
  - the availability of the processing capacity of each device,
  - The availability of the memory in devices, and
  - The bandwidth capacity of the participating devices.
- Solving the multi-objective optimization produces the rank of every device from faster to slower.
- Fed-MOODS adaptively involve devices in Federated learning.

Proposed System Model



Problem Formulations

- Multi-objective formulations for device selection

- Maximize available processing capacity

Processors utilization

$$D_u = \frac{1}{c}\sum_{i=1}^{c}(1-p^a)$$

$$D_c = (1-D_{cu})(\%)$$

$$D_g = (1-D_{gu})(\%)$$

$$\max_{i=1}^{N} D_i^{pa} = \frac{1}{2}(D_g + D_c)$$

$$\text{s.t. } 0 \le D_g \le 100, 0 \le D_C \le 100$$

[26] A. Yadin, *Computer Systems Architecture*. CRC Press, 2016.

- Maximize available memory

$$D_i^{MR} = B \times \sum_{l=1}^{L} MR_l \times Byte$$

$$D_i^{AMR} = D_i^{TM} - D_i^{MR}$$

$$\max_{i=1}^{N} D_i^{AMR}$$

$$\text{s.t. } \frac{D_i^{TM}}{D_i^{MR}} \ge 1, 0 \le D_i^{TM} \le 100$$

$$0 \le D_i^{MR} \le 100$$

https://cs231n.github.io/convolutional-networks/#case

- Maximize available bandwidth

$$D_i^{RNB} = \frac{D_i^{TD} * (100/D_i^{DR})}{(D_i^{RWT})}$$

$$\max_{i=1}^{N} D_i^{RNB}$$

$$\text{s.t. } D_i^{RWT} \ge 1, D_i^{DR} \ge 100, D_i^{TD} \ge 0$$

https://www.ibm.com/docs/en/tsmfsan/7.1.0?topic=SSSQZW_7.1.0/com.ibm.itsm.srv.doc/t_repl_est_bw.html

Problem Formulations

❖Federated learning objective

$$\mathcal{L}^{D_i}(w^{D_i}) = \frac{1}{m}\sum_{j=1}^{m} l(w_j, x_j^{D_i})$$

$$\mathcal{L}(w_\tau) \equiv \mathcal{L}_{n'}(w_{n'}) = \frac{1}{n'}\sum_{i=1}^{n'} \mathcal{L}^{D_i}(w^{D_i})$$

$$n' \leftarrow \min(n + \langle \tau, k \rangle, N)$$

Fed-MOODS

Phase 1 (Device rank)

1. Server collects meta-data to compute available processing capacity, memory, and bandwidth from N number of total devices.
2. Compute objective function of (1) available processing capacity, (2) available memory, (3) available bandwidth for each device.
3. Maximize these objective functions using Multi-objective optimization.
4. Rank each device based on the pareto fronts.

Phase 2

1. Select n' devices from pareto front.
2. Learn global model collaboratively.
3. Add another set of devices from the pareto fronts and learn global model.
4. Continue step 2 and 3 until model converges.

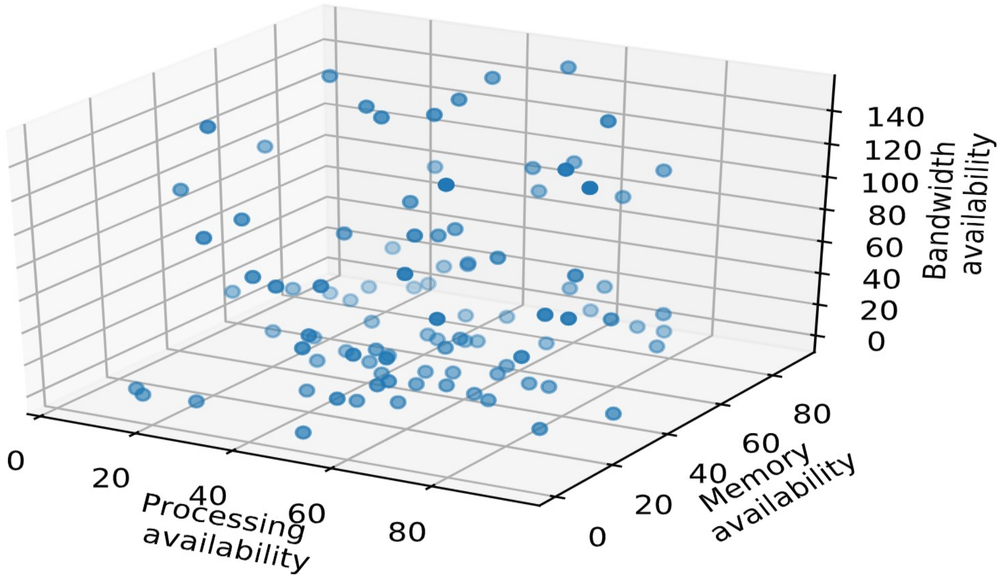
Experiments

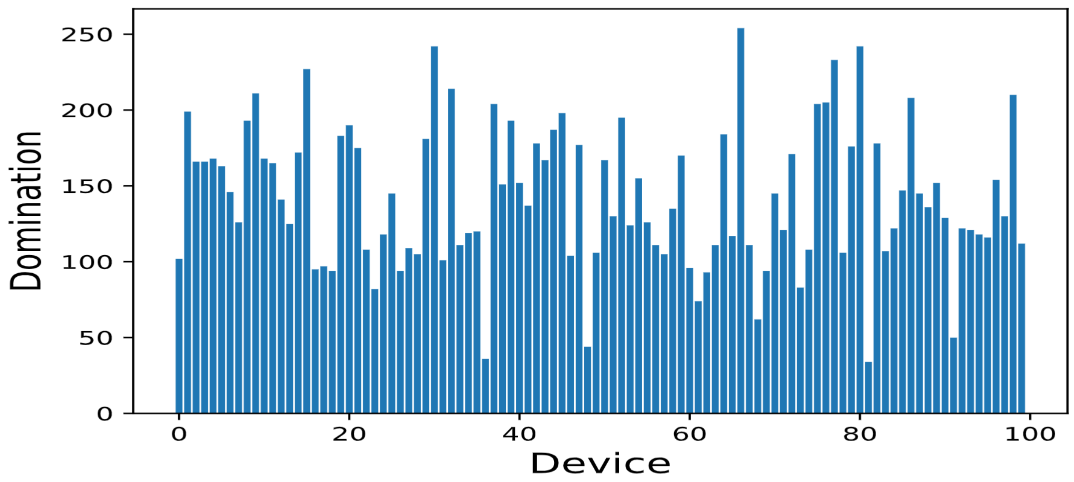
❖Simulation setup

Parameter(s)	Value	Description
local devices	100	Devices for a local update of the model
Server	1	For performing multi-objective optimization, model aggregation
Federated algorithm	2	Fed-Avg [14], Fed-Prox [6]
local device's participation	Adaptive and random	Adaptive participation of devices for Fed-MOODS, by random, frequency of participation is 10%
Dataset	IID and non-IID	IID and non-IID division of MNIST, CIFAR-10, and FMNIST dataset
Local iteration	Maximum 10	Number of local iteration at each device for each global iteration.
Global iteration	Maximum 100, 500	100 global iterations for learning on MNIST and FMNIST dataset. 500 global iterations for learning model on CIFAR-10 dataset.
Presence of stragglers	10%, 50%, 70%, 90%	Presence of stragglers in each global iteration for different experiments.
Training network	3	Three Convolutional Neural Network (CNN) having two hidden layers for training on MNIST, CIFAR-10, and FMNIST dataset respectively.
Optimizer	1	Stochastic Gradient Descent (SGD)
Performance metrics	2	Test accuracy, F1-score

Experiments

- ❖Phase 1
  - ❖Maximize three objective functions for all available devices.
  - ❖We have used NSGA-II, as a multi objective optimization algorithm.
  - ❖Calculate which device has highest domination.

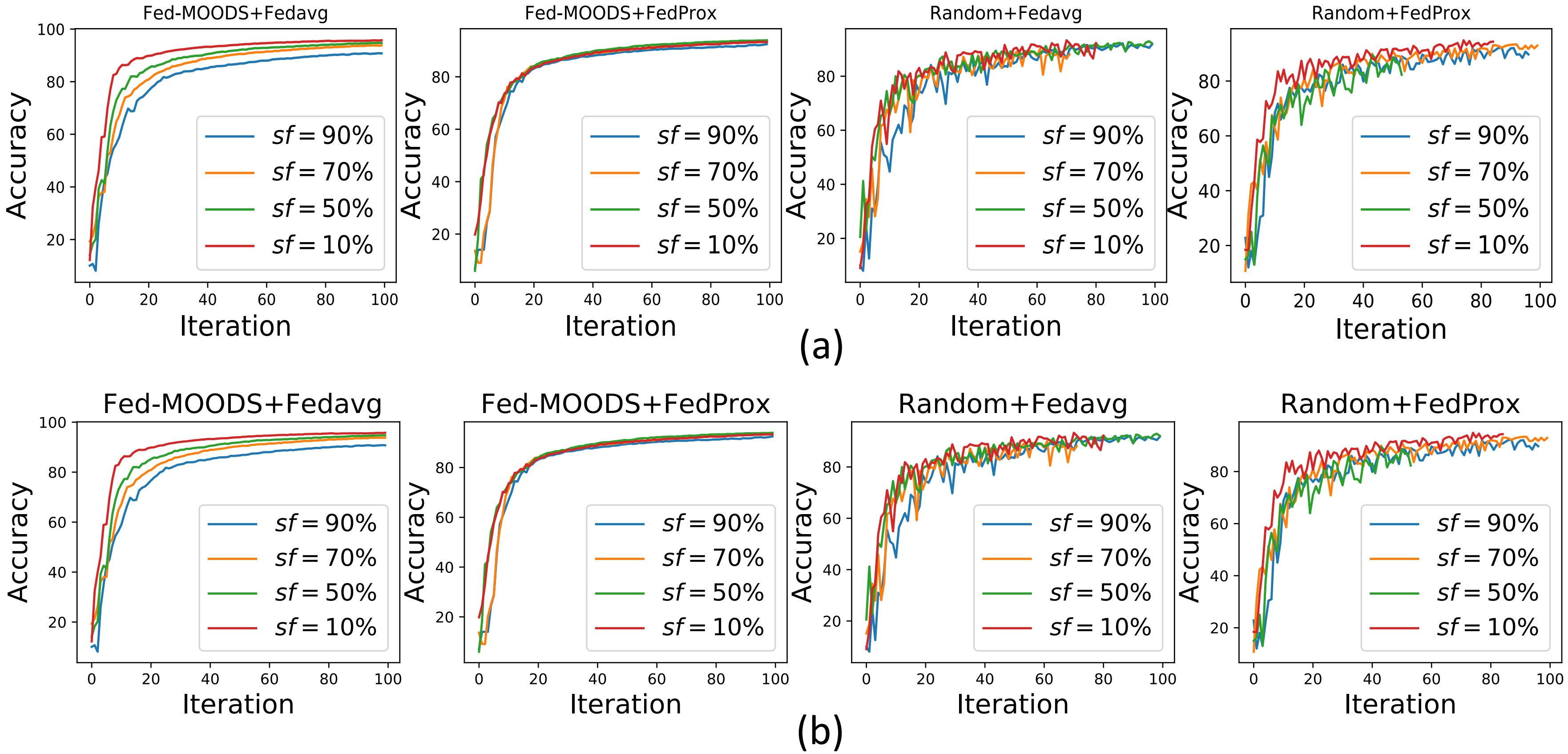




Experiments

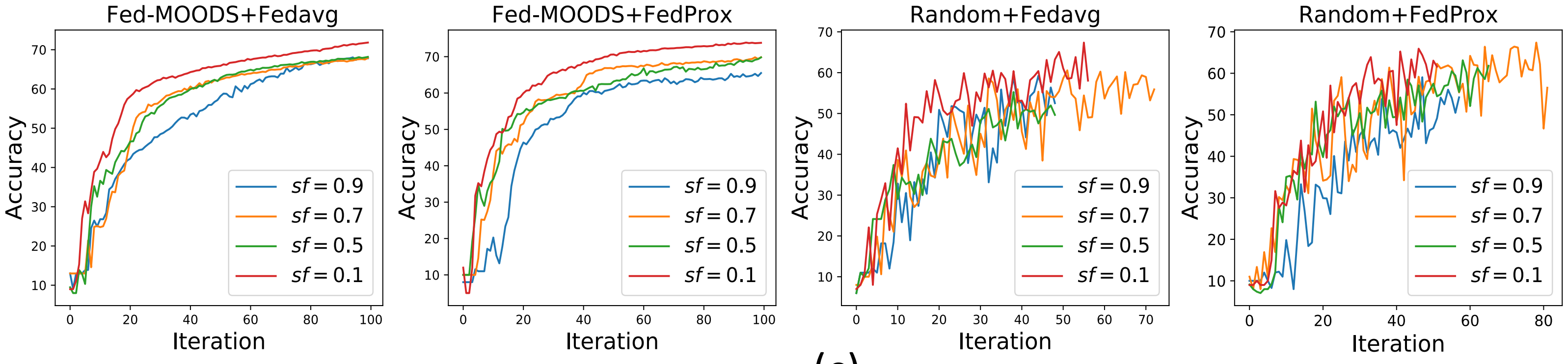
- Phase 2
  - Adaptive selection of devices based on the ranking of the devices.
- Results and analysis
  - Convergence comparison
  - Performance evaluation
  - Wall clock time analysis
  - Adaptiveness level analysis
  - Device fairness analysis

Results (Convergence comparison)

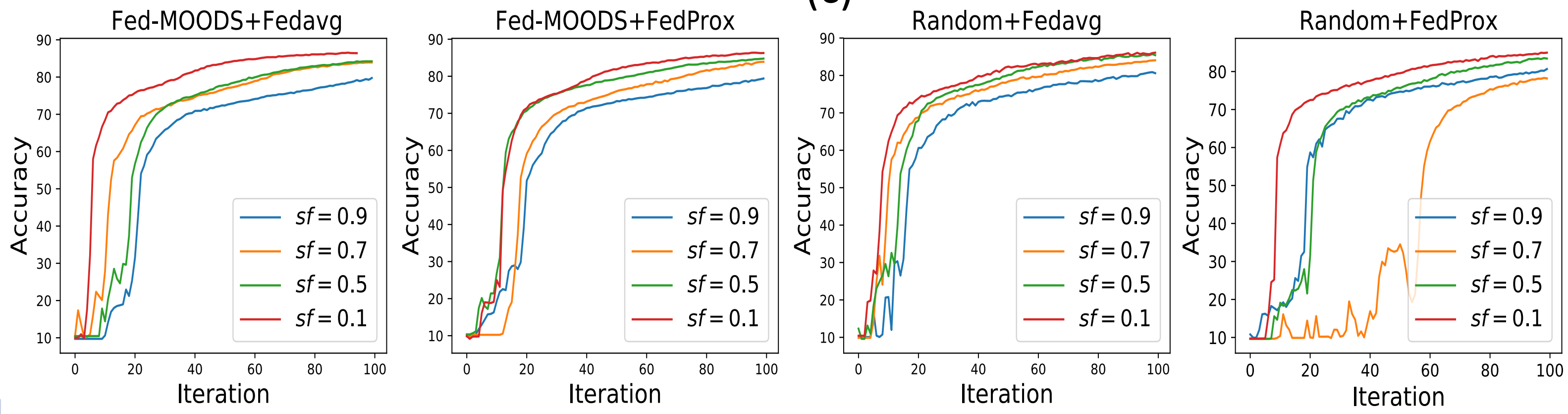


Convergence comparison of Fed-MOODS and benchmark models with random device participation across (a) MNIST-IID, (b) MNIST-non-IID, with different straggler fractions (sf).

Results (Convergence comparison)



(c)



(d)

Convergence comparison of Fed-MOODS and benchmark models with random device participation across (c) FMNIST IID, (d) FMNIST non-IID, with different straggler fractions (sf).

- ❖ Observation and Inference
  - ❖ Convergence of Fed-MOODS and random device selection are equivalent for the IID dataset.
  - ❖ For non-IID datasets, Fed-MOODS performs better than Random device selection.
  - ❖ Less effect of randomness is present.
  - ❖ Fed-MOODS + FedProx converges quicker than others.

Results (Performance comparison)

❖F1 score

PERFORMANCE (F1-SCORE) COMPARISON BETWEEN FED-MOODS AND BENCHMARK MODELS WITH RANDOM DEVICE PARTICIPATION IN PRESENCE OF 90% STRAGGLERS. ♥ AND ♢ DENOTE *involving stragglers* AND *not involving stragglers*, RESPECTIVELY.

Dataset	Fed-MOODS + FedAvg	Fed-MOODS + FedProx	Random device selection + FedAvg ♥	Random device selection + FedAvg ♢	Fed-MOODS + + FedAvg ♢
MNIST IID	94.7	93.5	94.00	96.28	**97.00**
CIFAR-10 IID	48.65	**52.92**	49.51	49.67	51.79
FMNIST IID	78.66	78.48	80.48	79.01	78.19
MNIST non-IID	93.41	**94.27**	93.00	NA	NA
CIFAR-10 non-IID	**49.33**	48.79	9.37	NA	NA
FMNIST non-IID	63.12	**65**	50.25	NA	NA

❖Observation and Inference

- ❖The performance of Fed-MOODS + FedProx is better than others in most cases.

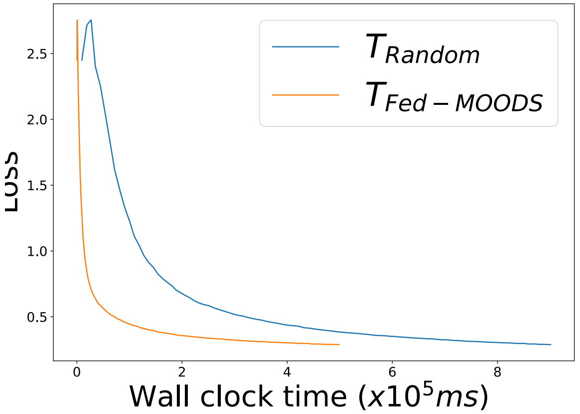
❖Test accuracy

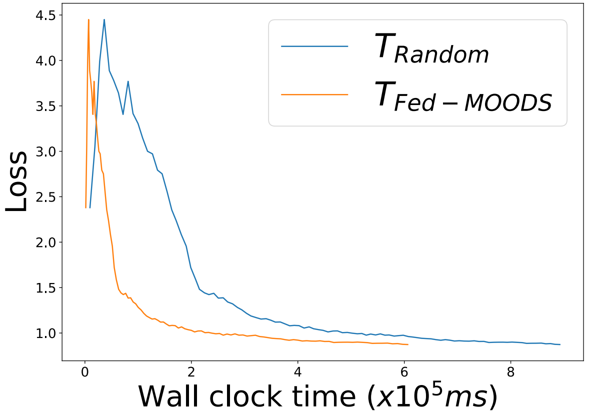
Dataset		SF %	Fed-MOODS + FedAvg	Fed-MOODS + FedProx	Random + FedAvg	Random + FedProx
MNIST	IID	90	**97.2**	96.31	97.2	96.89
		70	97.54	97.49	97.61	97.5
		50	**97.94**	97.76	97.74	97.61
		10	98.11	**98.39**	98.05	98.11
	Non-IID	90	92.31	91.93	92.04	. 91.47
		70	**93.91**	92.79	89.18	93.43
		50	**94.69**	93.47	93.05	89.61
		10	**95.74**	93.59	93.17	93.86
CIFAR-10	IID	90	**53.43**	50.20	49.15	48.86
		70	46.3	47.15	48.62	47.17
		50	43.59	**49.42**	46.25	48.9
		10	46.71	**47.33**	45.48	44.72
	Non-IID	90	49.23	**49.55**	15.84	10
		70	**48.75**	47.68	33.99	29.75
		50	**46.56**	45.93	24.98	38.44
		10	45.86	**47.81**	33.75	34.0
FMNIST	IID	90	78.66	78.48	80.48	79.44
		70	82.63	82.81	83.04	77.63
		50	83.32	83.89	85.17	82.59
		10	**85.39**	85.22	84.44	84.68
	Non-IID	90	63.22	**65.33**	50.26	58.18
		70	**67.16**	65.54	56.92	64.07
		50	70.0	**70.97**	55.56	61.81
		10	**71.76**	67.58	58.18	59.26

Results (Wall-clock time comparison)

Total and average wall clock time comparison between Fed-MOODS and benchmark model with random device selection at presence of 90% stragglers on non-IID data.

Datasets	Random Device selection		Fed-MOODS	
	$T_{Random}(ms)$	$\bar{T}_{Random}(ms)$	$T_{Fed-MOODS}(ms)$	$\bar{T}_{Fed-MOODS}(ms)$
MNIST	9×10^5	9×10^3	$\mathbf{4.9 \times 10^5}$	$\mathbf{4.9 \times 10^3}$
FMNIST	8.9×10^5	8.9×10^3	$\mathbf{6 \times 10^5}$	$\mathbf{6 \times 10^3}$



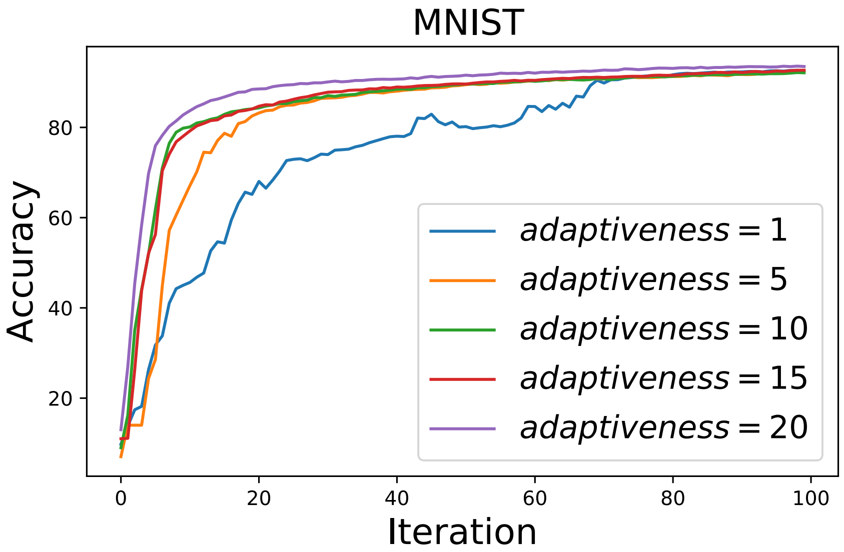


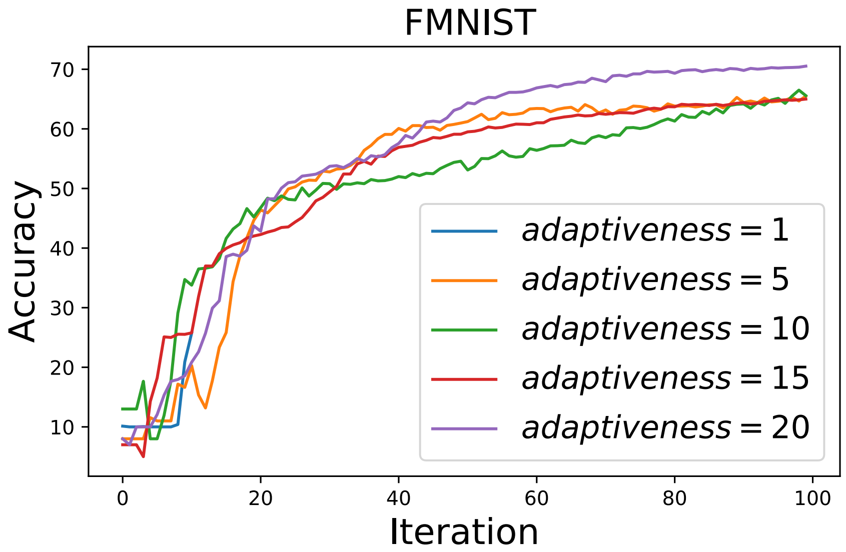
- ❖ Observation and Inference
  - ❖ Fed-MOODS converging before Random device selection.
  - ❖ Fed-MOODS is 1.8× and 1.48× faster than the benchmark model (FedAvg) with random device participation on the MNIST and FMNIST non-IID dataset, respectively.

Training loss vs wall-clock time comparison of Fed-MOODS and benchmark model with random device participation in presence of 90%, stragglers respectively on MNIST non-IID (left) and FMNIST non-IID (right) data.

Results (Adaptiveness)

90% devices are stragglers.





- ❖ Observation and Inference
  - ❖ Increase in adaptivenss produce better convergence.
  - ❖ Larger adaptiveness means involving stragglers in more iteration.

Analysis (Device fairness)

- ❖Does Fed-MOODS maintain fairness?
  - ❖Yes.
  - ❖Every device is getting a chance to contribute to the learning process.
  - ❖The probability of appearance(PoA) of non-stragglers is high and for stragglers is low. Where as, In random device selection, every device has the same PoA.

Conclusion

- We proposed Fed-MOODS, a multi-objective optimization-based adaptive device selection method to minimize the effect of stragglers in federated learning.
- We formulated the available processing capacity, available memory, and available bandwidth of every device as a multi-objective optimization problem and by solving it we generate the rank of devices.
- The algorithm adaptively selects devices for training based on the ranking.
- We found that Fed-MOODS is 1.8x and 1.48x faster than the random device selection method for MNIST and FMNIST dataset.

Future work

❖Theoretical proof of the convergence of the adaptive method.

❖Theoretical value for the adaptiveness level.

❖ Device profiling is enough ? Or few more to consider ?

Acknowledgement

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

References

[1] S. Banerjee et. al., “Multi-diseases classification from chest-x-ray: A federated deep learning approach,” in *Australasian Joint Conference on Artificial Intelligence*. Springer, pp. 3–15,2020

[2] Y. Wang et. al., “Accelerated Training via Device Similarity in Federated Learning,” in *Proc. of the 4th In Workshop on Edge Sys., Analytics and Networking*, pp. 31–36, 2021.

[3] C. Yang et. al, “Characterizing Impacts of Heterogeneity in Federated Learning upon Large-Scale Smartphone Data,” in *Proc. of the Web Conference*, pp. 935–946, 2021.

[4] V. Smith et. al, “Federated Multi-task Learning,” *arXiv:1705.10467*, 2017.

[5] P.Kairouzet.al, “Advances and Open Problems in Federated Learning,” *arXiv:1912.04977*, 2019.

[6] T. Li et. al, “Federated optimization in heterogeneous networks,” *arXiv:1812.06127*, 2018.

[7] X. Li et. al, “On the convergence of fedavg on non-iid data,” *arXiv:1907.02189*, 2019.

[8] A. Reisizadeh et. al., “Straggler-Resilient Federated Learning: Leveraging the Interplay Between Statistical

Accuracy and System Heterogeneity,” *arXiv:2012.14453*, 2020.

[9]B. McMahan et. al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Artificial int. and stats.* PMLR, pp. 1273–1282, 2017.

[10] S. Banerjee et. al., “Fed-fis: a novel information-theoretic federated feature selection for learning stability,” in *ICONIP,*. Springer, pp. 480-487, 2021

[11] X.Lianet.al,“AsynchronousDecentralizedParallelStochasticGradient Descent,” in *Int. Conf. on Machine Learning*. PMLR, pp. 3043– 3052, 2018

[12] S. Stich, “Local SGD Converges Fast and Communicates Little,” *arXiv:1805.09767*, 2018.

[13] C. Xie et. al, “Asynchronous Federated Optimization,” *arXiv:1903.03934*, 2019.

References

[14] A. Mitra et. al, “Achieving Linear Convergence in Federated Learn- ing under Objective and Systems Heterogeneity,” *arXiv preprint arXiv:2102.07053*, 2021.

[15] H. Shi et. al., “Towards Federated Learning with Attention Transfer to Mitigate System and Data Heterogeneity of Clients,” in *Proc. of the 4th Int. Workshop on Edge Sys., Analytics and Networking*, pp. 61–66., 2021

[16] G. Canonaco et. al., “Adaptive federated learning in presence of concept drift,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–7, 2021.

[17] L.Liet.al.,“Fedsae: A novel self-adaptive federated learning framework in heterogeneous systems,” in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–10, 2021.

[18] S. Cui et. al., “Addressing Algorithmic Disparity and Performance Inconsistency in Federated Learning,” *arXiv:2108.08435*, 2021.

[19] Z. Hu et. al, “Fedmgda+: Federated learning meets multi-objective optimization,” *arXiv:2006.11489*, 2020.

[20] Y. Shi et. al., “A survey of fairness-aware federated learning,” *arXiv preprint arXiv:2111.01872*, 2021.

[21] P. Zhou et. al., “Loss tolerant federated learning,” *arXiv preprint arXiv:2105.03591*, 2021.

[22] T.Huanget.al.,“An efficiency-boosting client selection scheme for federated learning with fairness guarantee,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1552–1564, 2020.

[23] M. Mohri et. al., “Agnostic federated learning,” in *International Conference on Machine Learning*. PMLR, pp. 4615–4625, 2019

[24] T. Li et. al., “Fair resource allocation in federated learning,” *arXiv preprint arXiv:1905.10497*, 2019.

[25] L.Lyuet.al.,“Collaborative fairness in federated learning, Springer, pp. 189–204, 2020

[26] A. Yadin, *Computer Systems Architecture*. CRC Press, 2016.

[27] K. Deb et. al, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Trans. On evolutionary comp.*, vol. 6, no. 2, pp. 182– 197, 2002.

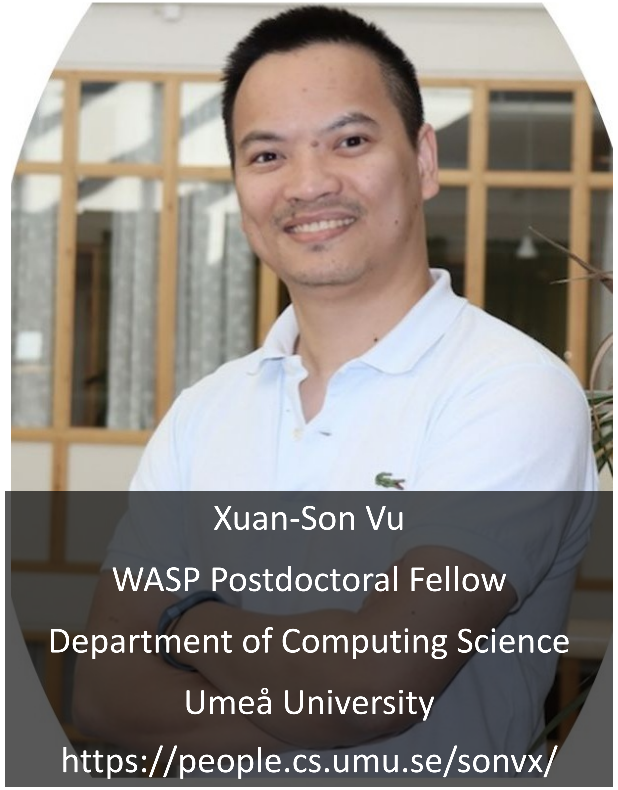
[28] L. Deng, “The mnist database of handwritten digit images for machine learning research [best of the web],” *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[29] A. Krizhevsky et. Al., “Learning multiple layers of features from tiny images,” 2009.

[30] H. X. et. Al. , Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017

Authors

Sourasekhar Banerjee
WASP PhD. Student
Department of Computing Science
Umeå University
https://www.umu.se/en/staff/sourasekhar-banerjee/



Xuan-Son Vu
WASP Postdoctoral Fellow
Department of Computing Science
Umeå University
https://people.cs.umu.se/sonvx/

Monowar Bhuyan
WASP Assistant Professor
Department of Computing Science
Umeå University
https://people.cs.umu.se/monowar/index.html

Thank You!

sourasb@cs.umu.se