Dual Adaptive Compression for Efficient Communication in Heterogeneous Federated Learning

Yingchi Mao[a,b], Zibo Wang[b], Chenxin Li[b], Jiakai Zhang[b], Shufang Xu[a,b], and Jie Wu[c]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources,
Hohai University, Nanjing, China

[b] School of Computer and Information, Hohai University, Nanjing, China

[c] Center for Networked Computing, Temple University, Philadelphia, USA

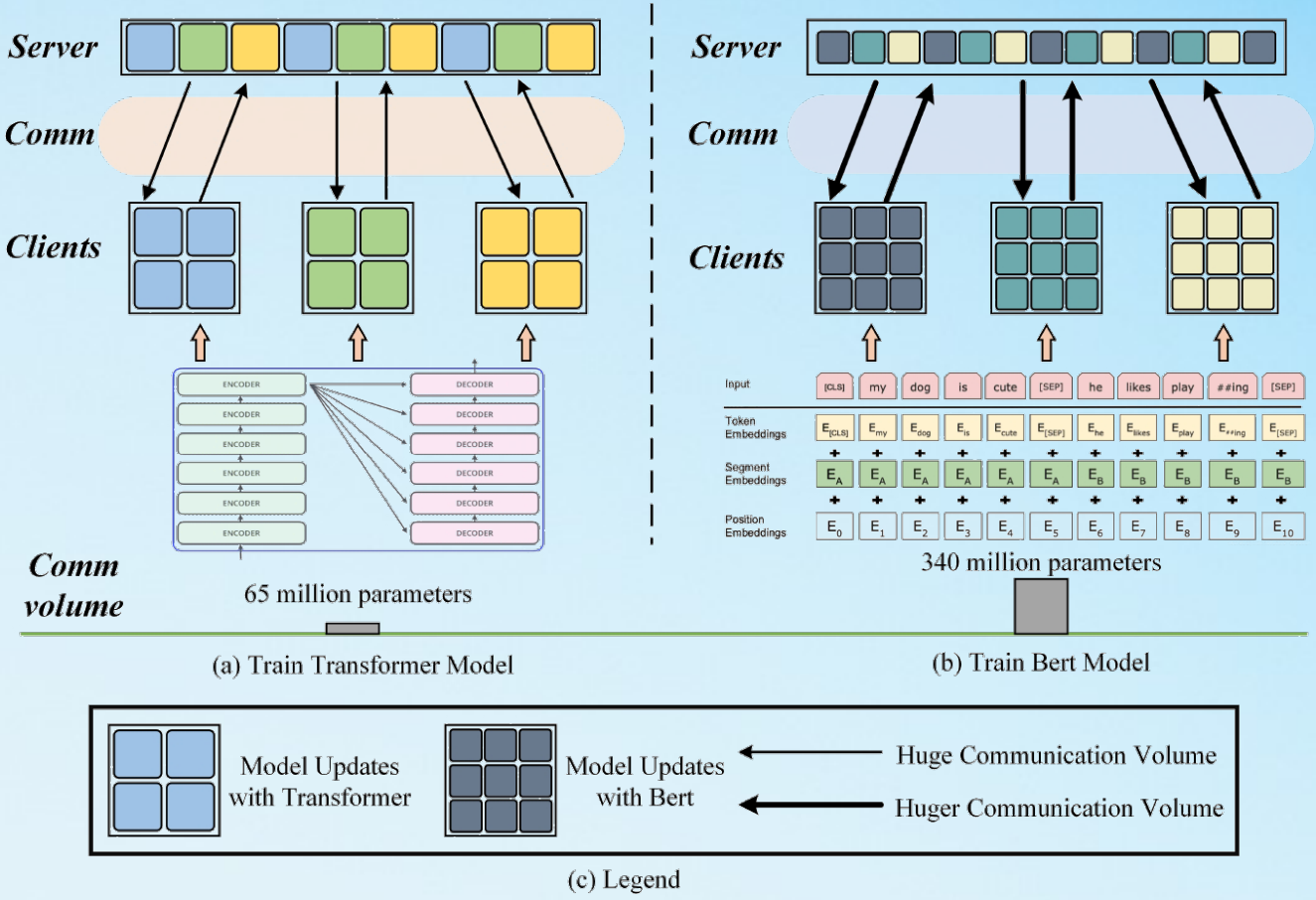
IEEE/ACM CCGRID
The 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing
May. 6-9, 2024

Content

Introduction

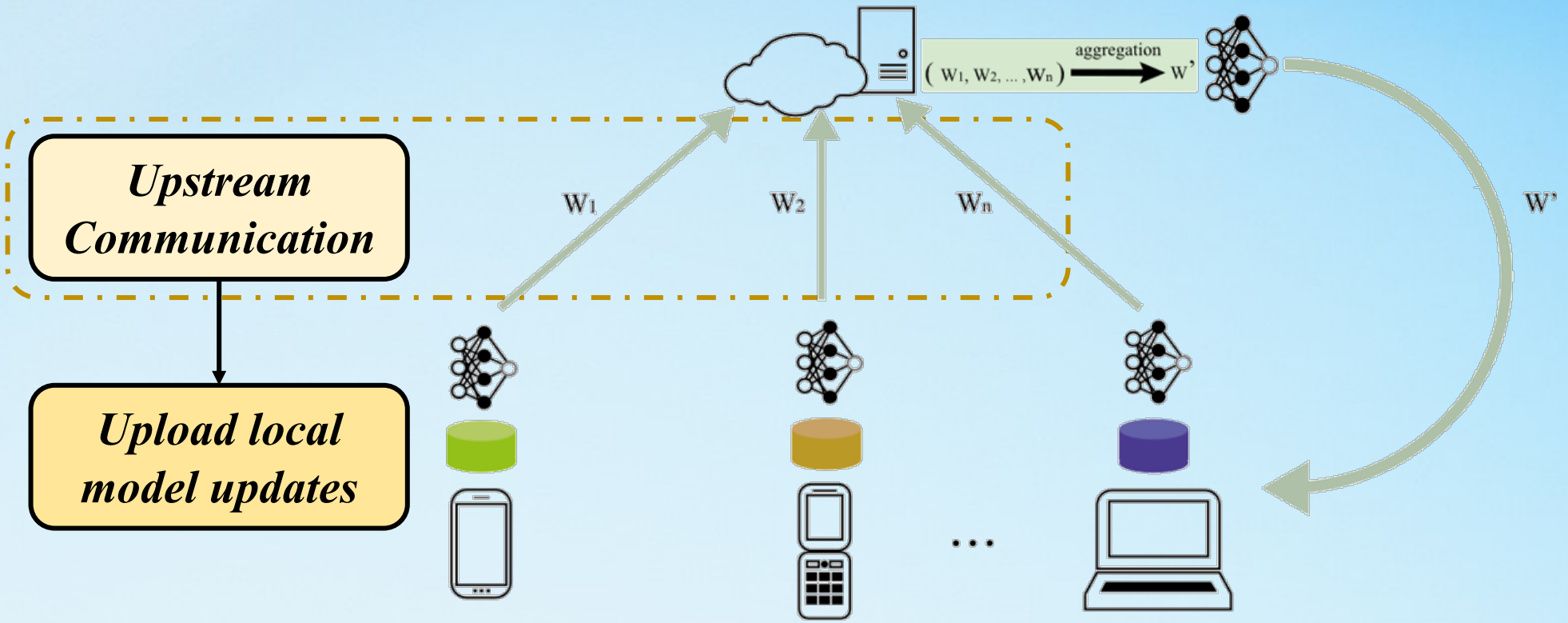
■ Background

In federated learning, the model updates transmitted between clients and the server lead to significant communication costs, especially when it involves large-scale models.



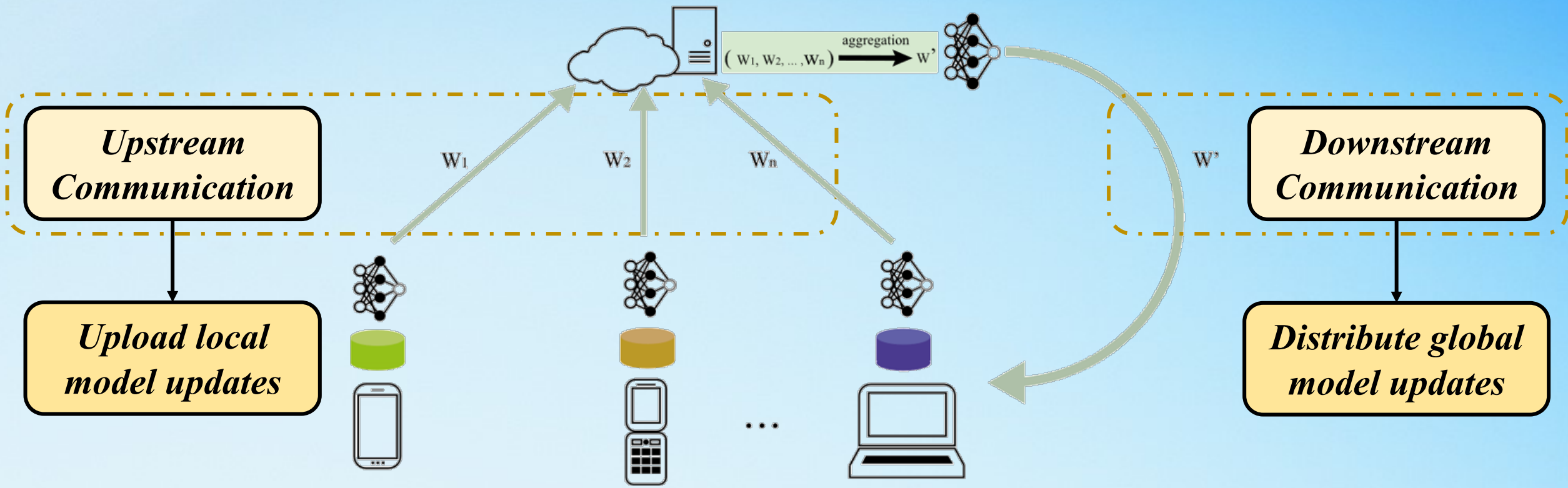
Introduction

- During the upstream communication of federated learning, clients upload their local model updates to the server, and then the server update the global model.



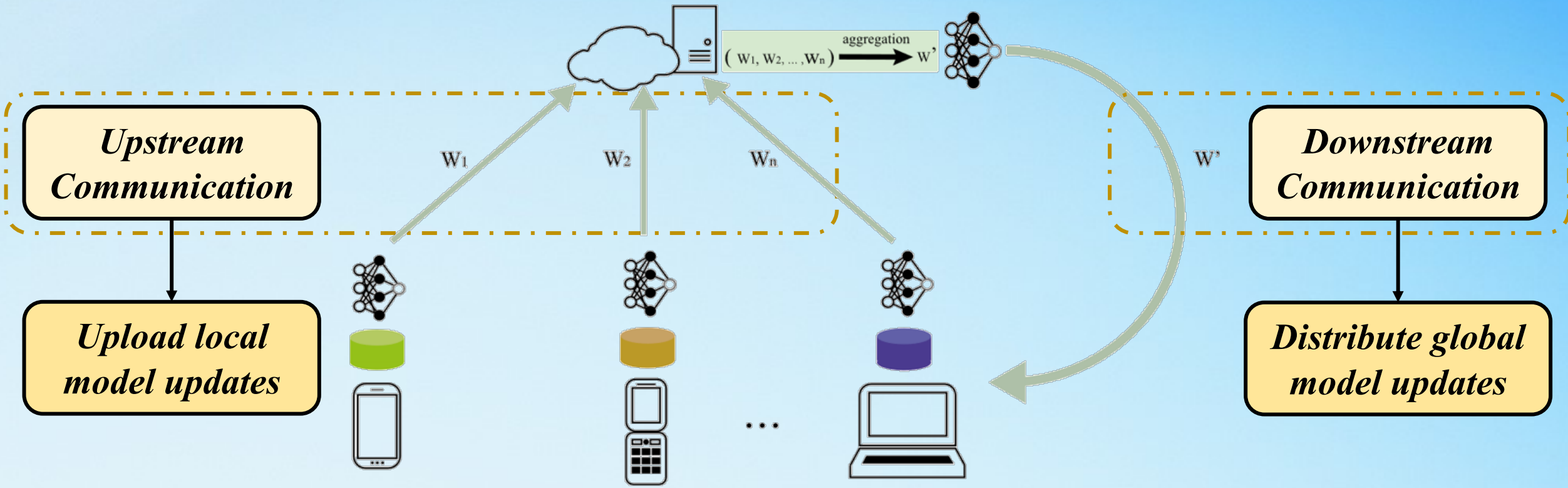
Introduction

- During the downstream communication, the server distributes the global model updates to initialize the global model on each client.



Introduction

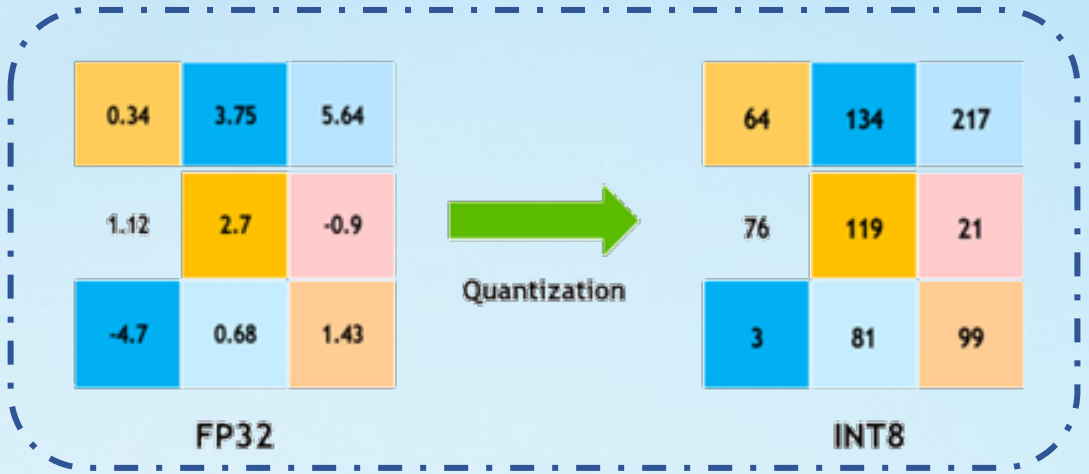
- We can compress the model updates transmitted between clients and the server to reduce the single round communication volume, there by reducing the communication costs in federated learning.



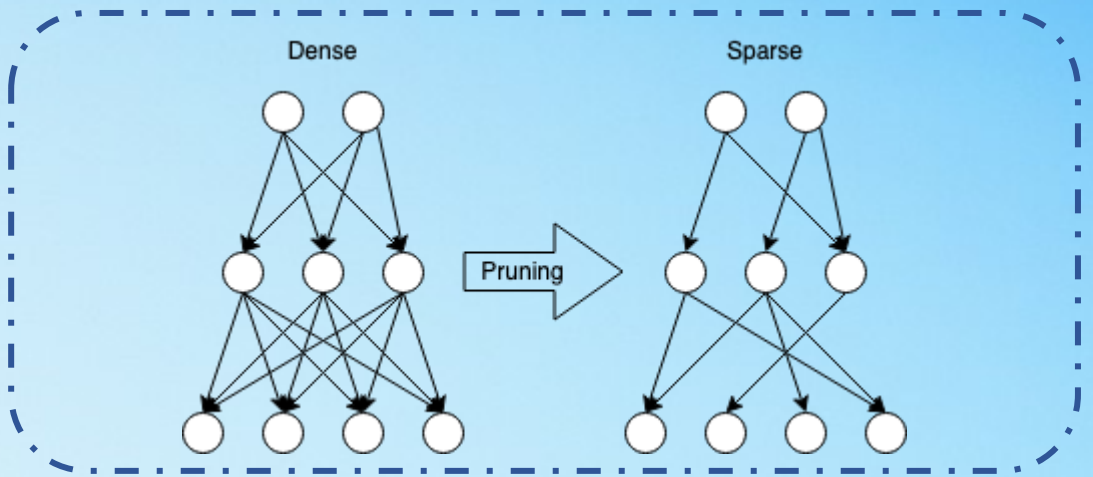
Compress the model updates → *Reduce the single round communication volume*

Introduction

- As typical model compression techniques, quantization and sparsification are commonly utilized in distributed machine learning for efficient communication.



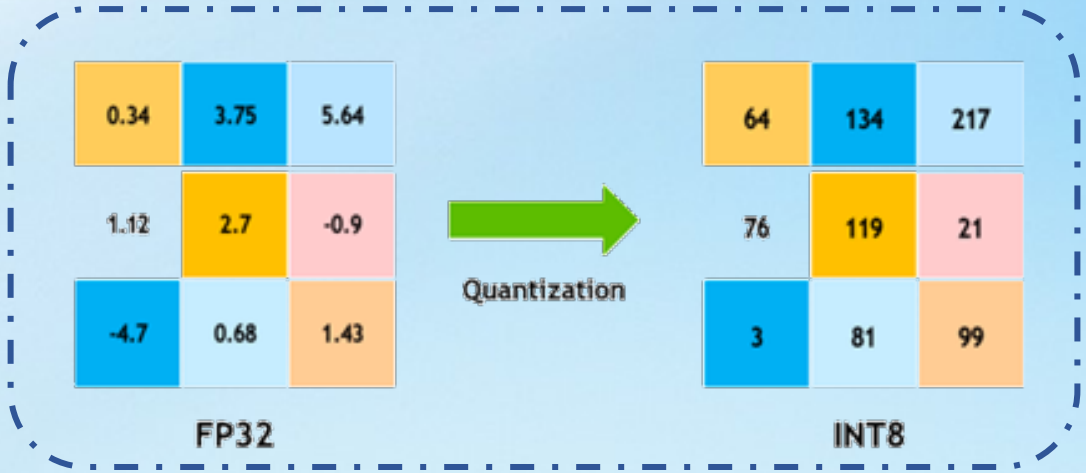
Quantization



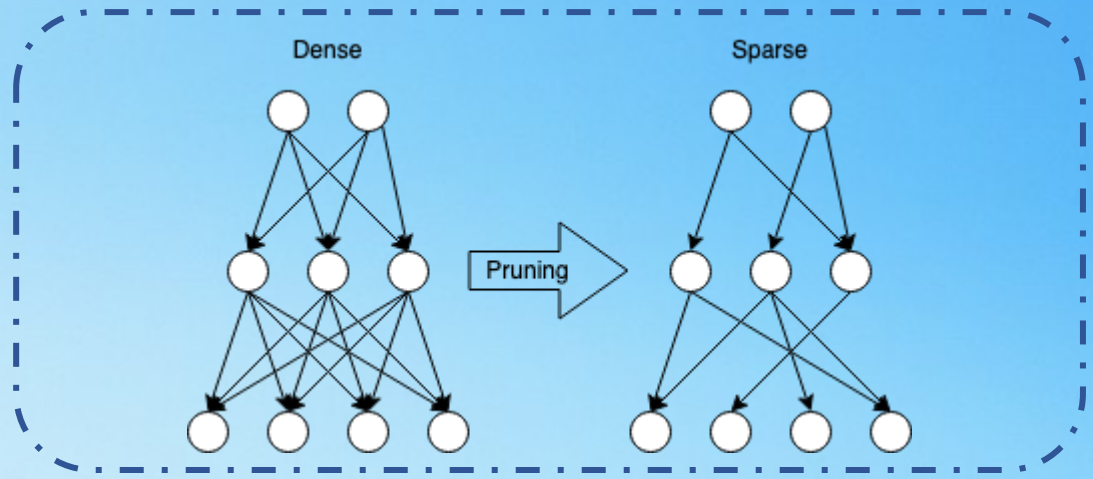
Sparsification

- However, quantization and sparsification cannot be directly employed in federated learning.

Introduction



Quantization



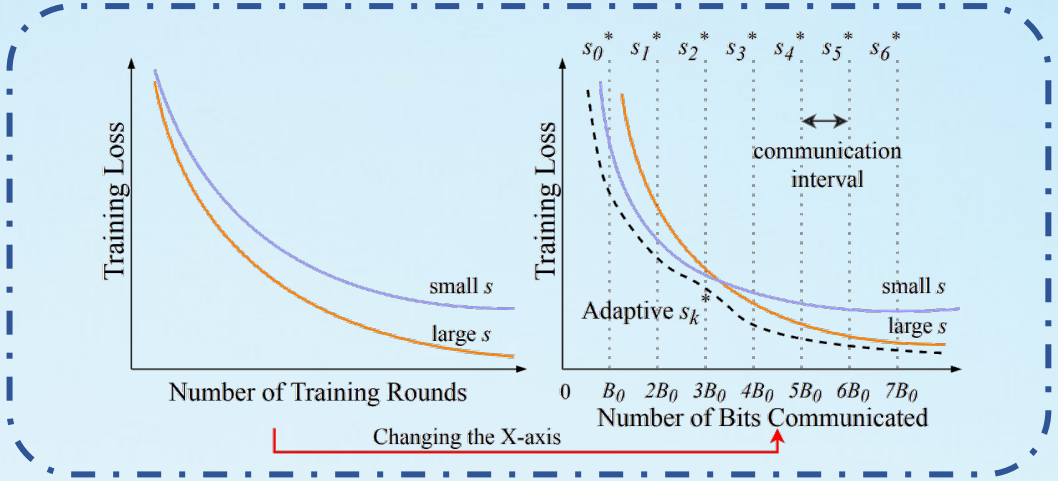
Sparsification

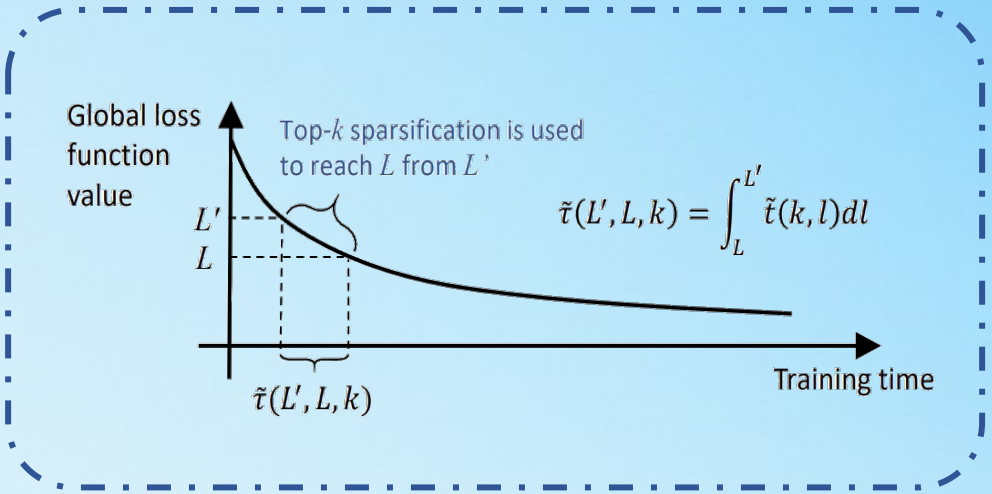
- Lossy compression in quantization or sparsification leads to a decline in model accuracy, it is challenging to strike a balance between communication efficiency and model accuracy.
- In heterogeneous federated learning, employing the same and fixed compression coefficients for all clients with different data distributions will exacerbate gradient conflict and gradient drift.

Related Work

■ One-way Adaptive compression methods

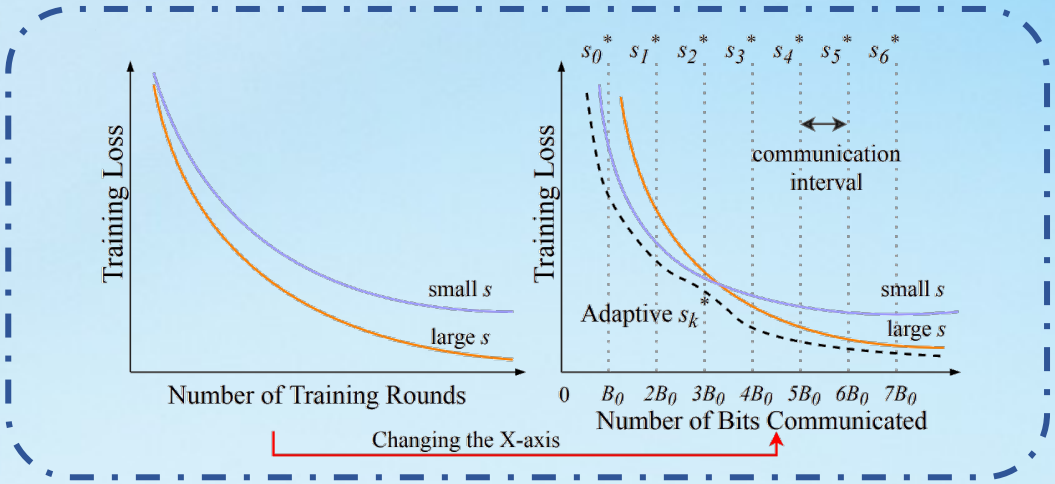
- In order to deploy suitable compression methods for heterogeneous federated learning, adaptive compression is introduced by related studies.

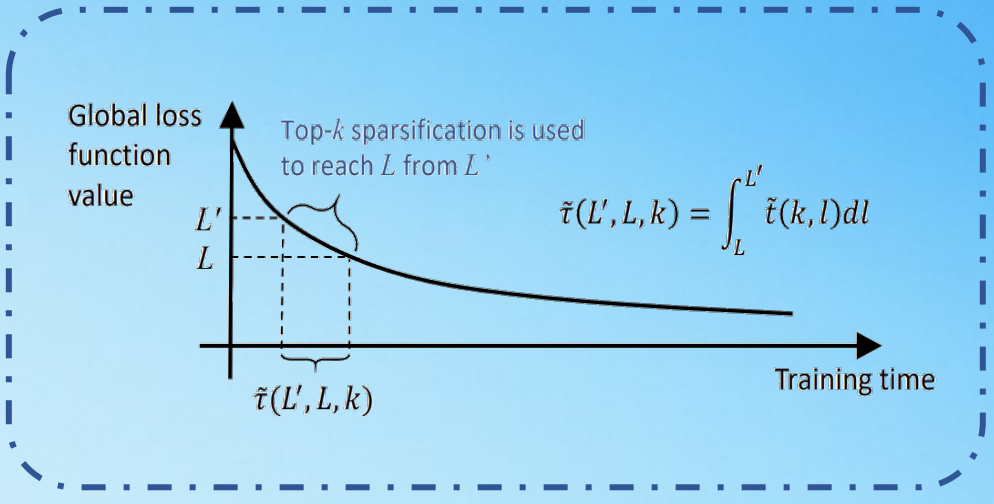




Related Work

■ One-way Adaptive compression methods





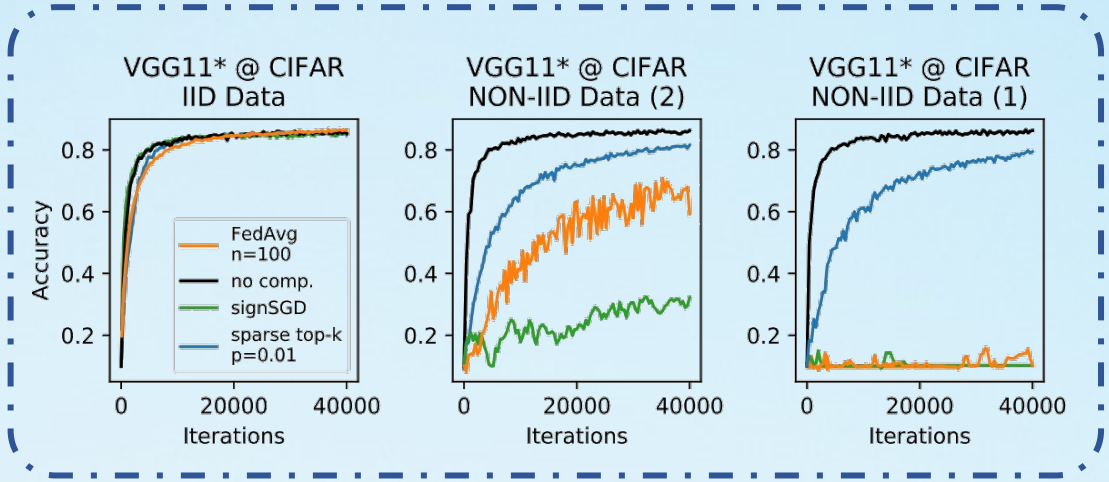
- AdaQuantFL: It utilizes the initial and current local loss to adjust the quantization coefficient, mitigating the problem of high error floor due to quantization, [D. Jhunjhunwala et al. ICASSP'2021]
- FAB-top-k: It predicts the fluctuation of the loss function to adjust the sparsity ratio, [P. Han et al. ICDCS'2020]

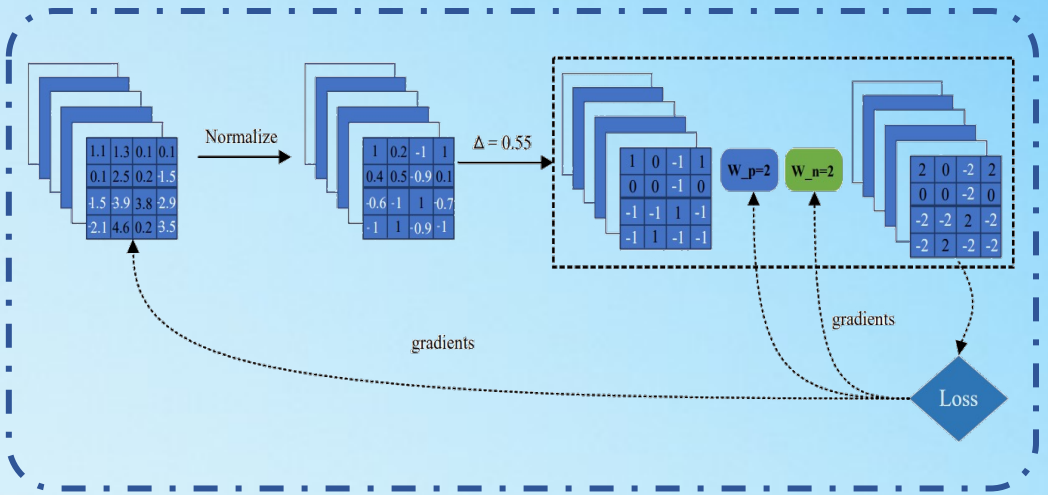
These methods only reduce the upstream communication volume, and the communication efficiency needs to be further optimized.

Related Work

■ Two-way compression methods

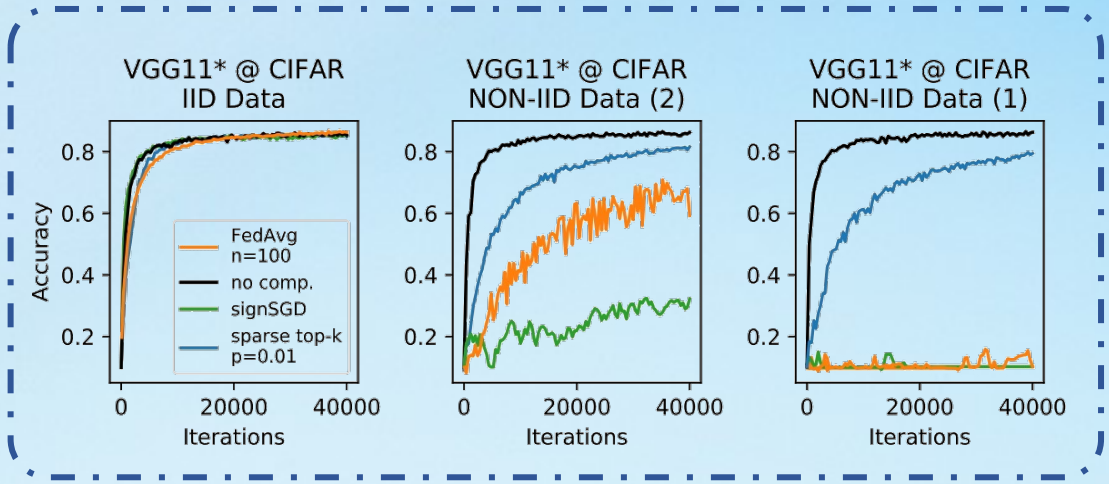
- To achieve this, various mechanisms are introduced in the relevant researches, thereby simultaneously compressing model updates transmitted between upstream and downstream.

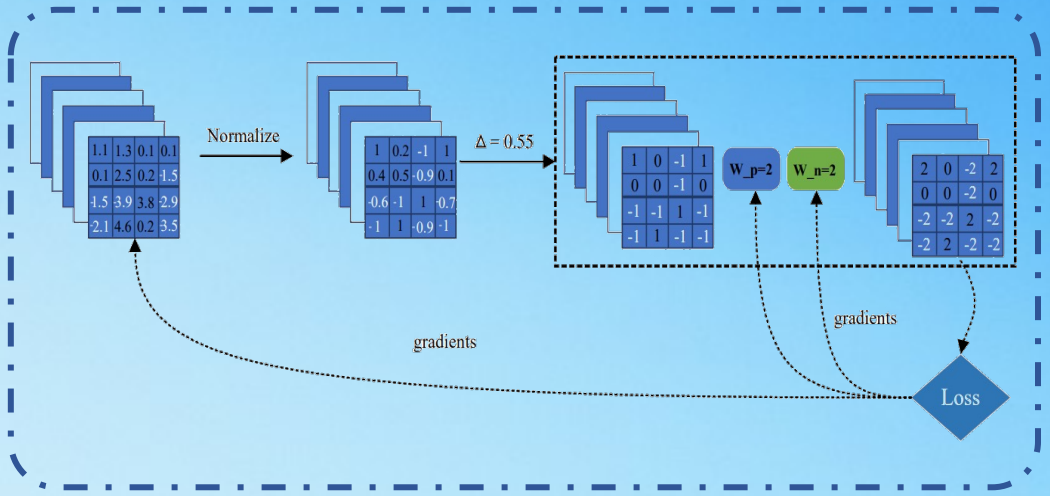




Related Work

■ Two-way compression methods



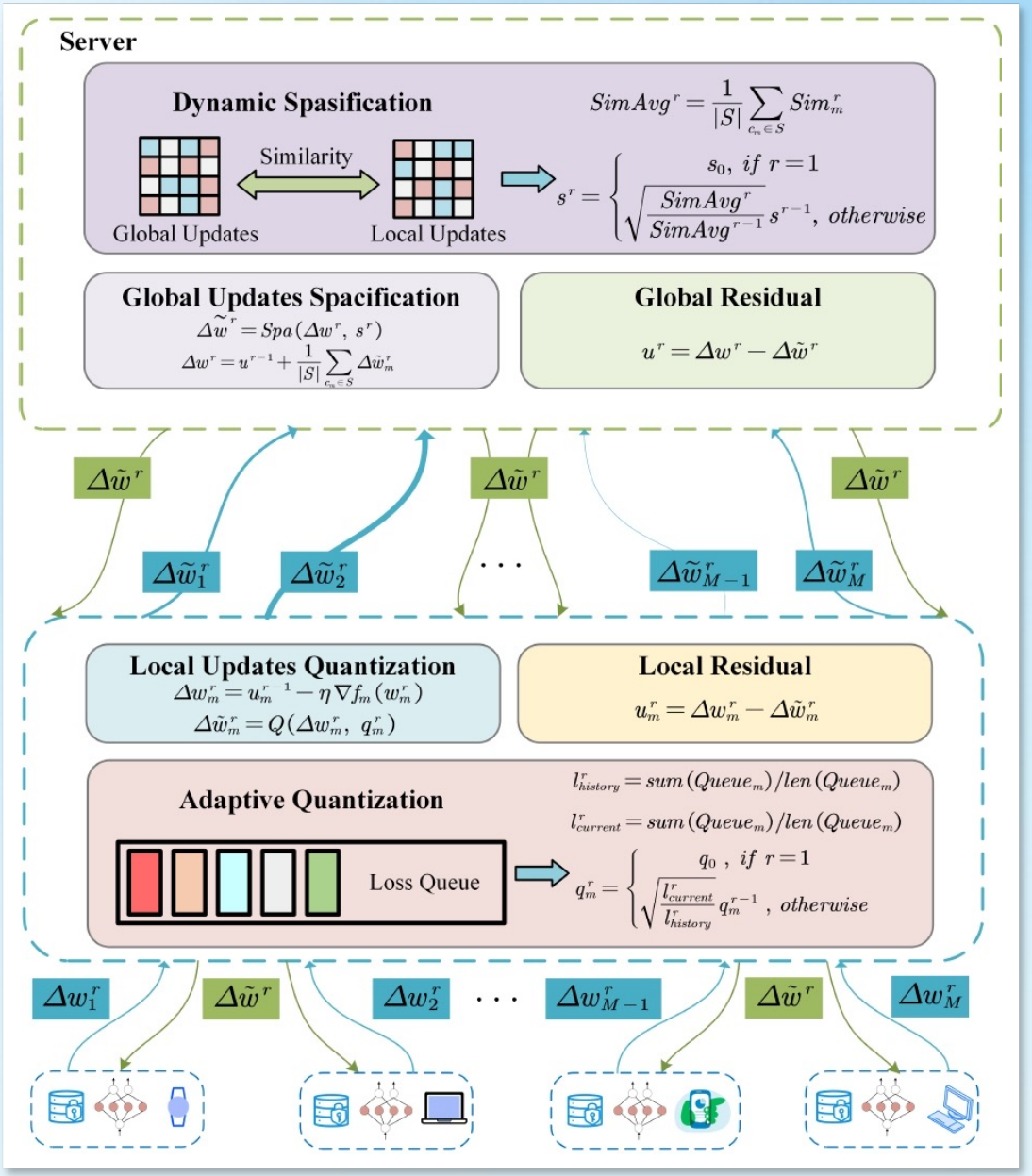


- STC: Combined with sparsification and ternary quantization, STC can extremely reduce the communication costs in federated learning, [F. Sattler et al. IEEE NNLS'2020]

- T-FedAvg: It quantizes the local updates and sparsifies global updates during training, achieving bidirectional compression for upstream and downstream communication, [J. Xu et al. IEEE NNLS'2022]

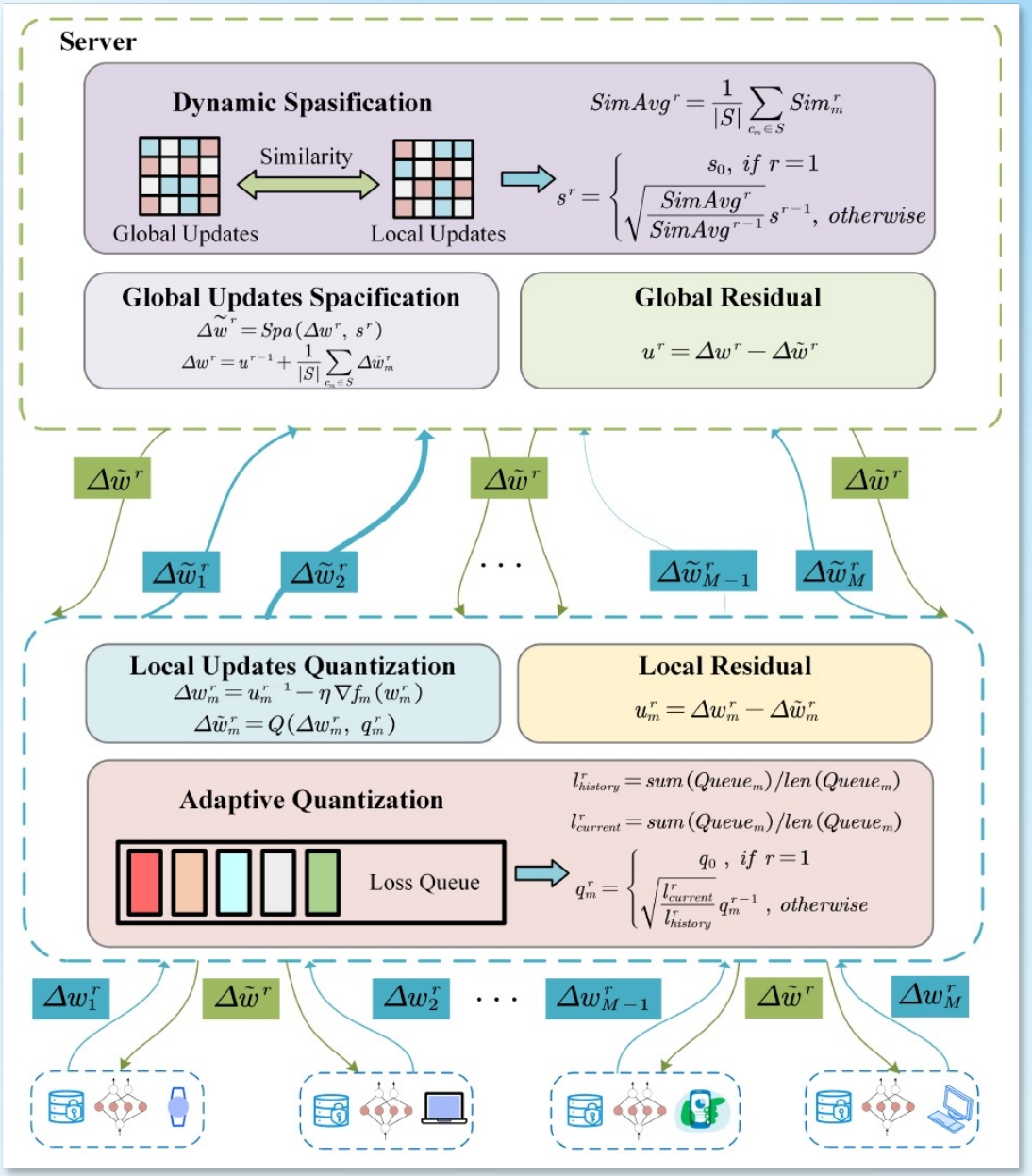
These methods adopt a fixed compression ratio, which cannot be directly employed in heterogeneous federated learning.

Approach

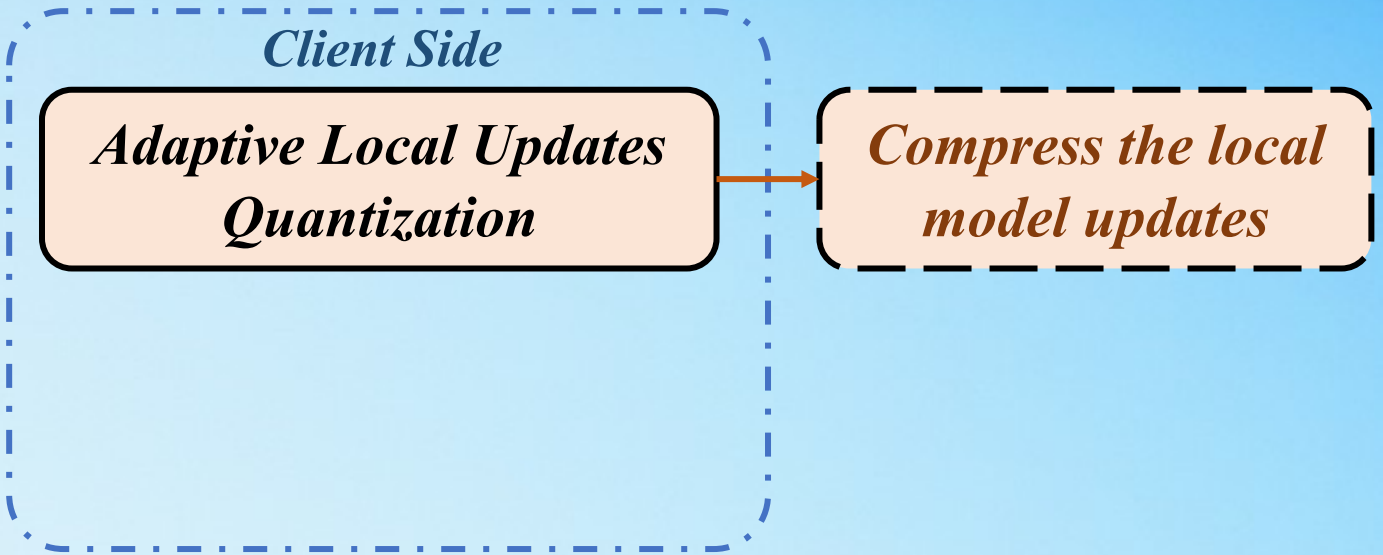


- To achieve dynamic bidirectional compression while balancing the model accuracy and communication efficiency in heterogeneous federated learning, we propose a **dual adaptive compression method**(***FedDAC***) in this study.

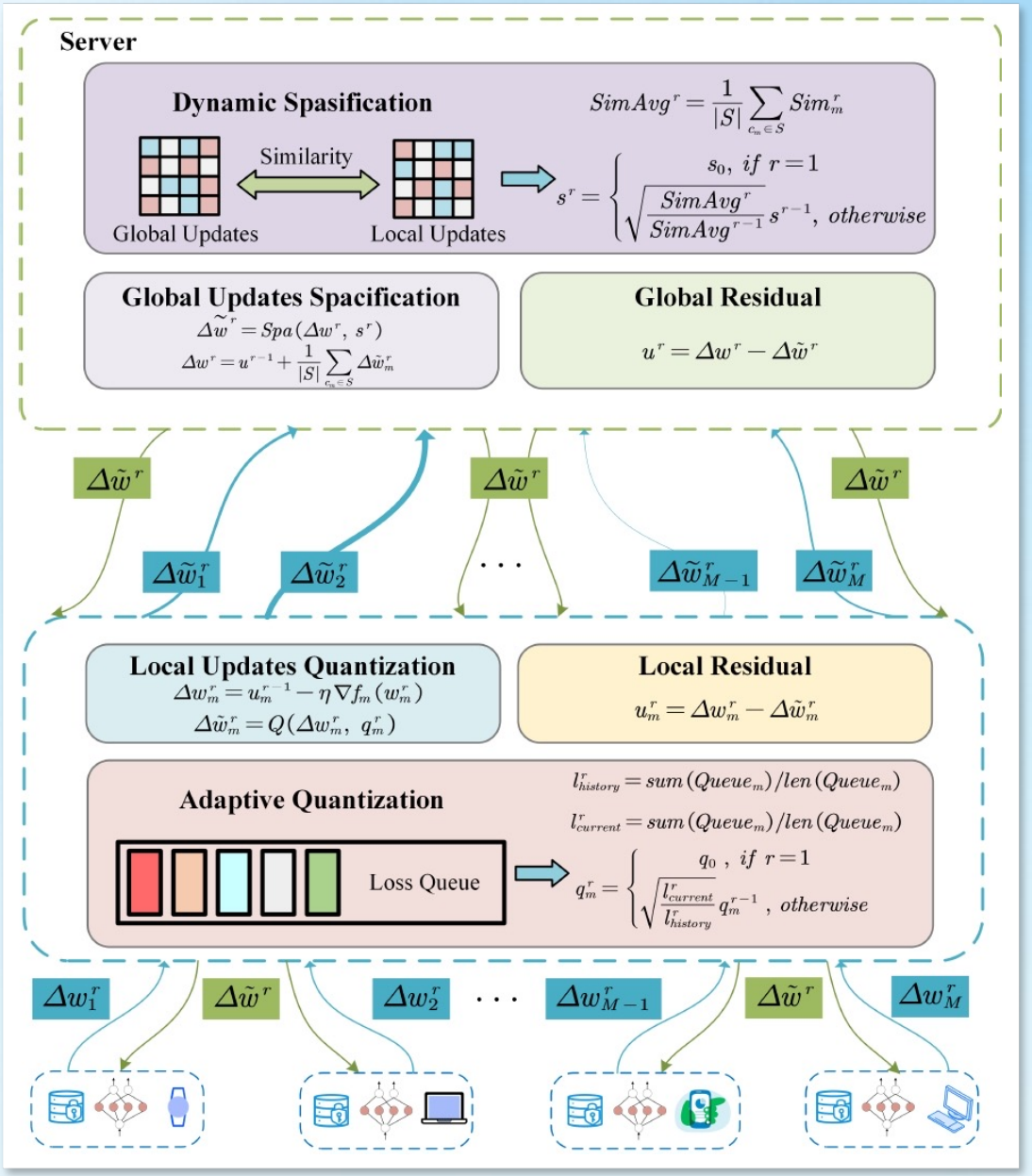
Approach



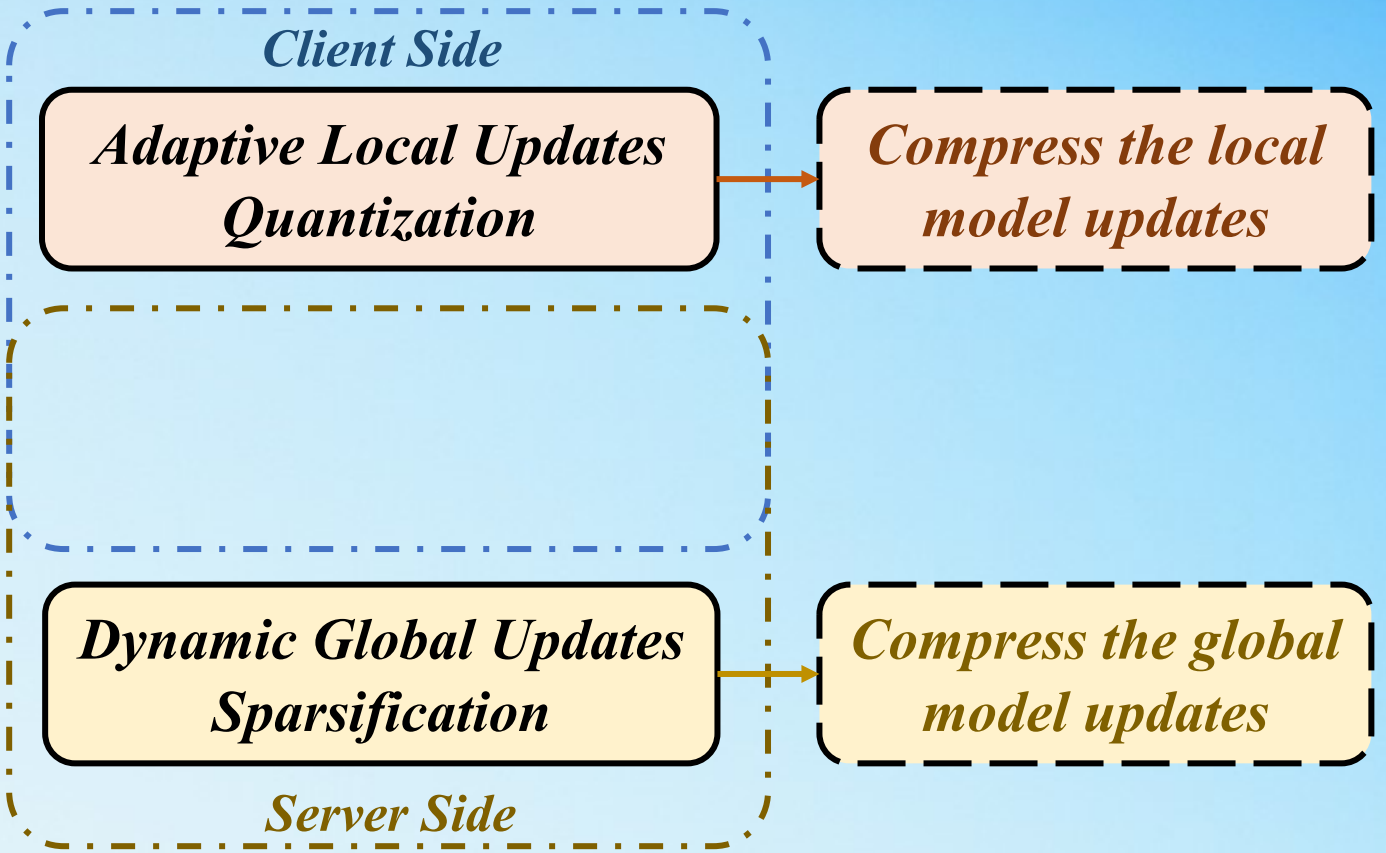
- Specifically, on the client side, we introduce the ***Adaptive Local Updates Quantization*** module to compress the local model updates.



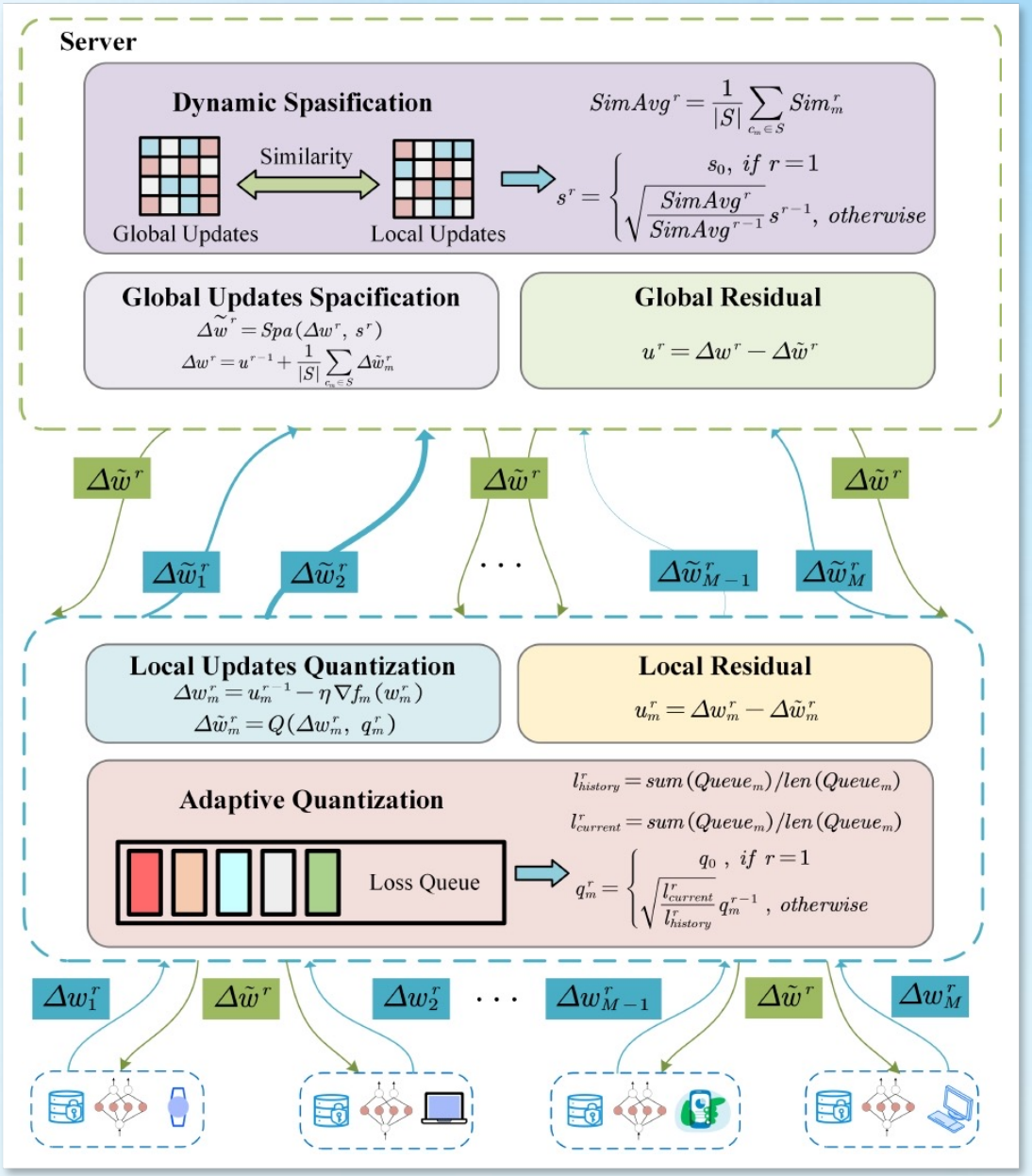
Approach



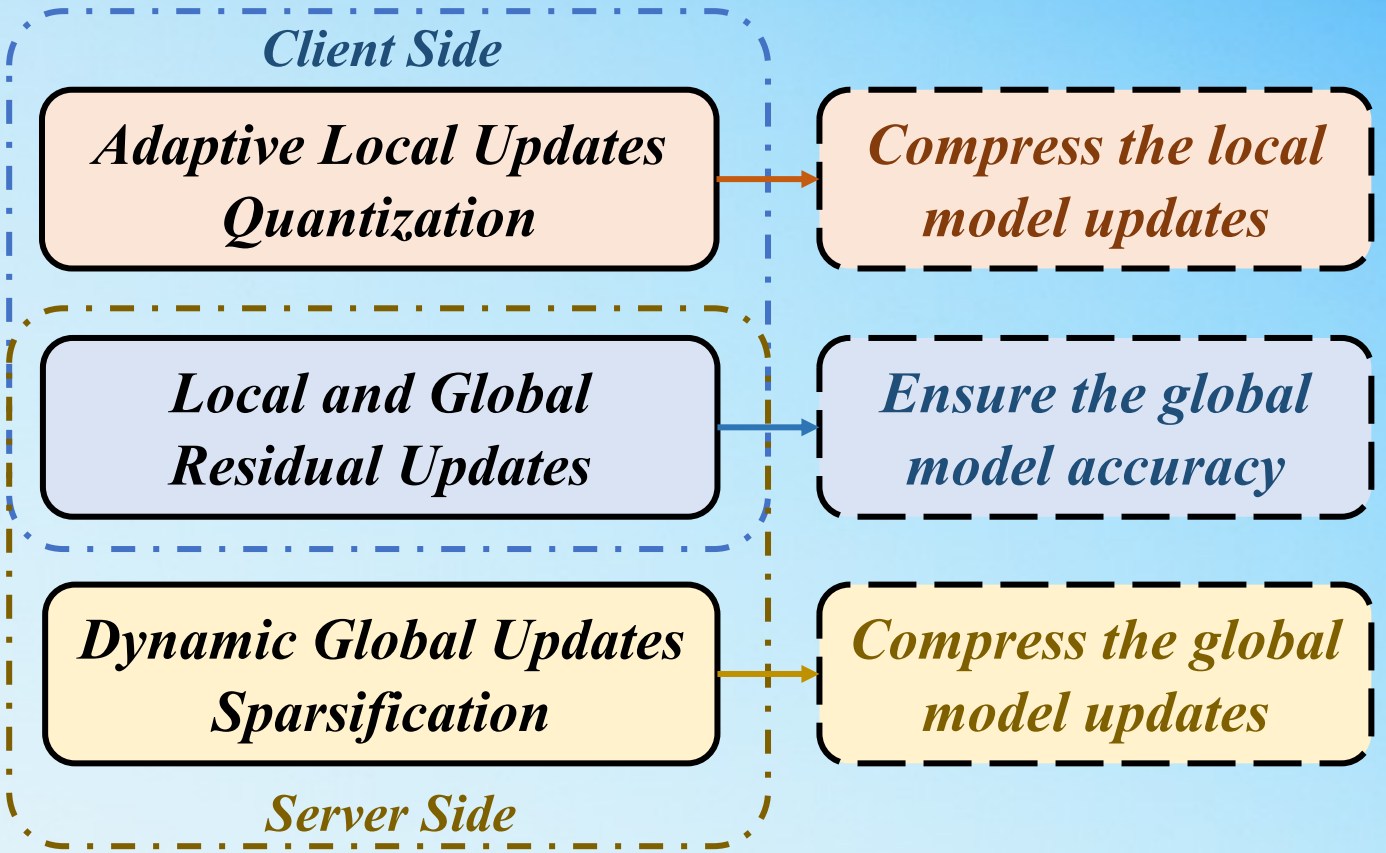
- On the server side, the ***Dynamic Global Updates Sparsification*** module is employed to compress the global model updates.



Approach



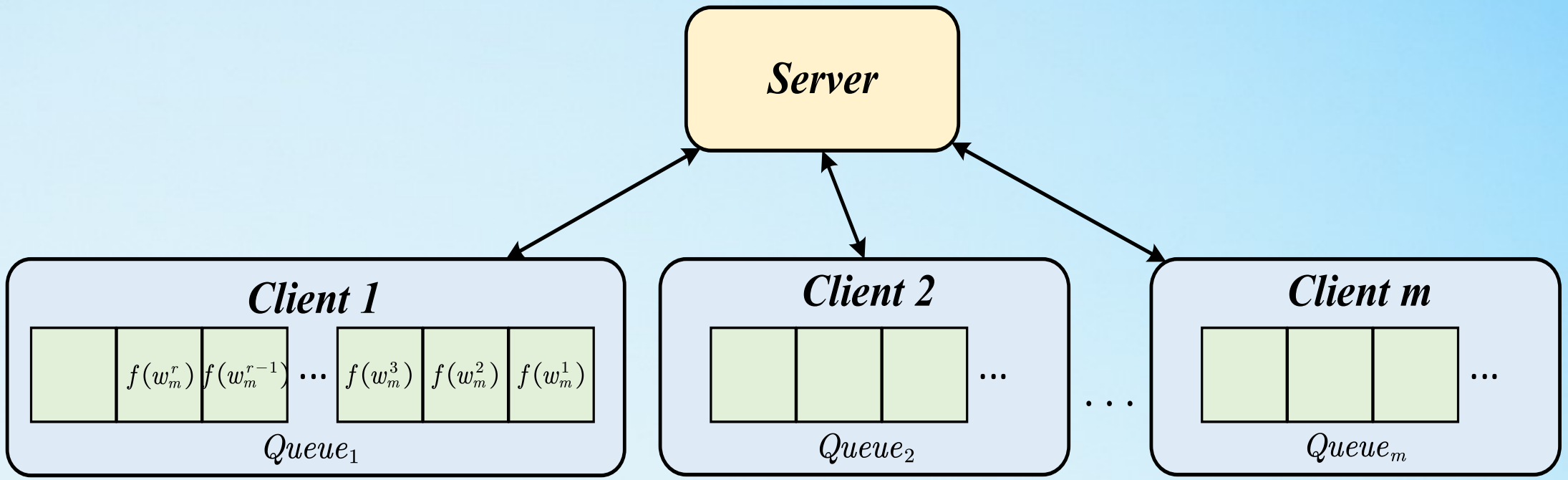
- In order to mitigate the degradation of global model accuracy caused by lossy compression, we introduce the ***Local and Global Residual Updates*** on both the client and the server side.



Approach

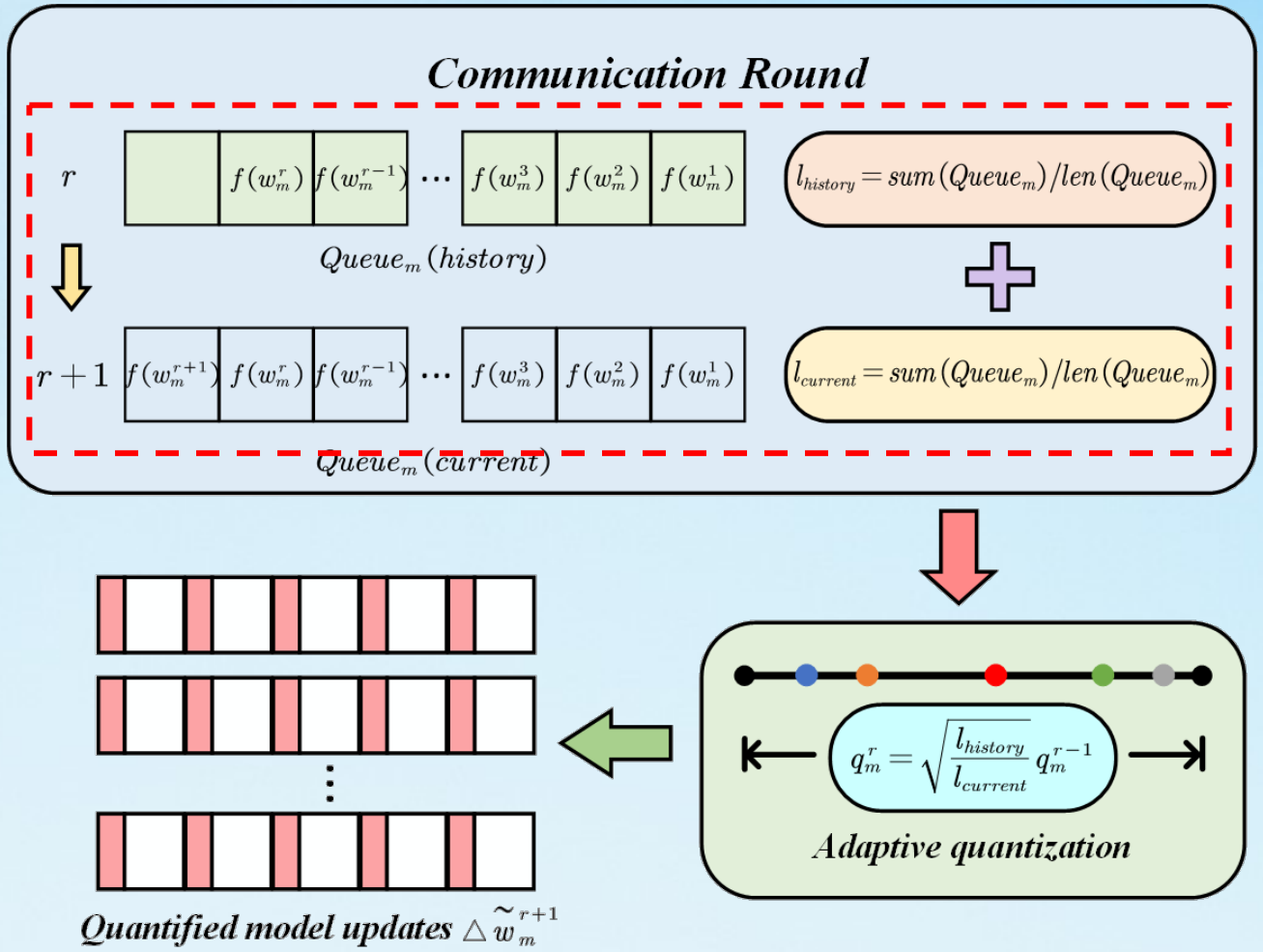
Adaptive Local Updates Quantization

- In the local computation phase, a loss queue of capacity μ is defined on each client, which is utilized to store the local loss of each client.
- With the introduction of the loss queue, the convergence trends within each client can be determined.



Approach

Adaptive Local Updates Quantization



- In the *r-th* iteration, the client *m* computes the historical average loss, and then adds its *r-th* loss to the loss queue, then computes the current average loss.

Historical/Current average loss

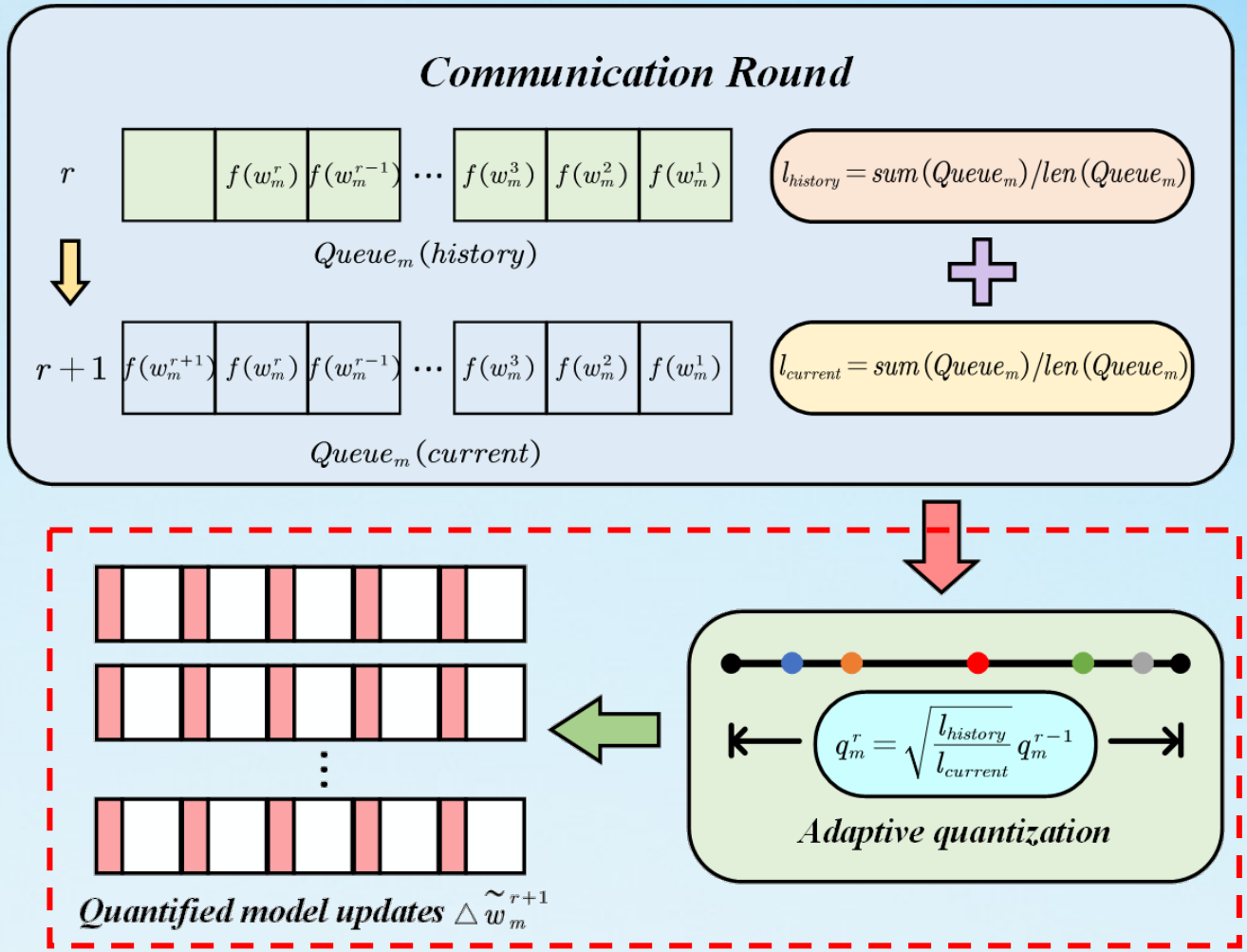
$$l_{history}^r = sum(Queue_m)/len(Queue_m)$$

$$enqueue(Queue_m,\ loss_m)$$

$$l_{current}^r = sum(Queue_m)/len(Queue_m)$$

Approach

Adaptive Local Updates Quantization



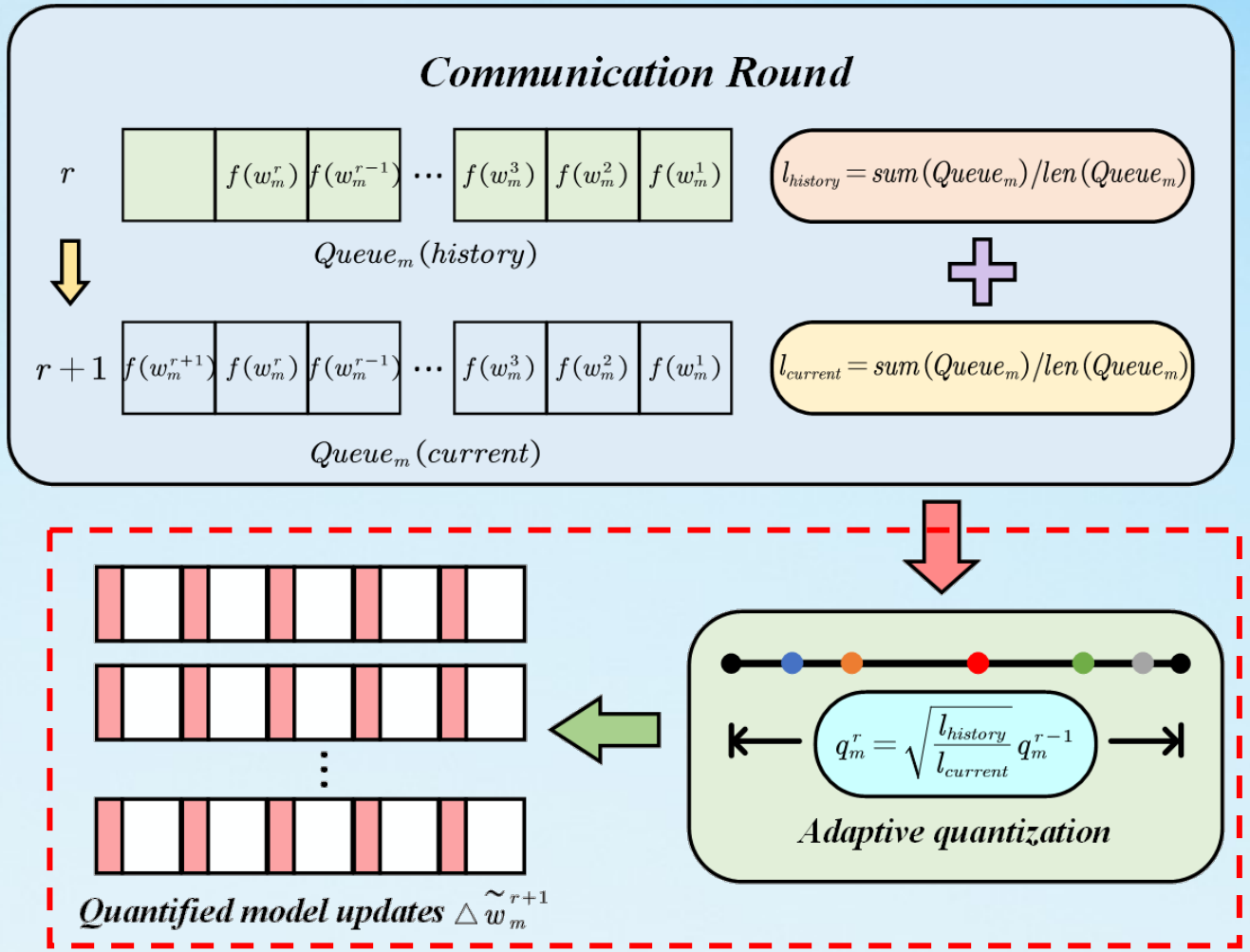
- The quantization coefficient of client *m* in the *r-th* iteration is determined based on the ratio of the current average loss and historical average loss.

Quantization coefficient calculation

$$q_m^r = \begin{cases} q_0 \ , \ if \ r = 1 \\ \sqrt{\frac{l_{current}^r}{l_{history}^r}}\, q_m^{r-1} \ , \ otherwise \end{cases}$$

Approach

Adaptive Local Updates Quantization



- The client *m* will quantize its local model updates based on the quantization coefficient.

Quantization coefficient calculation

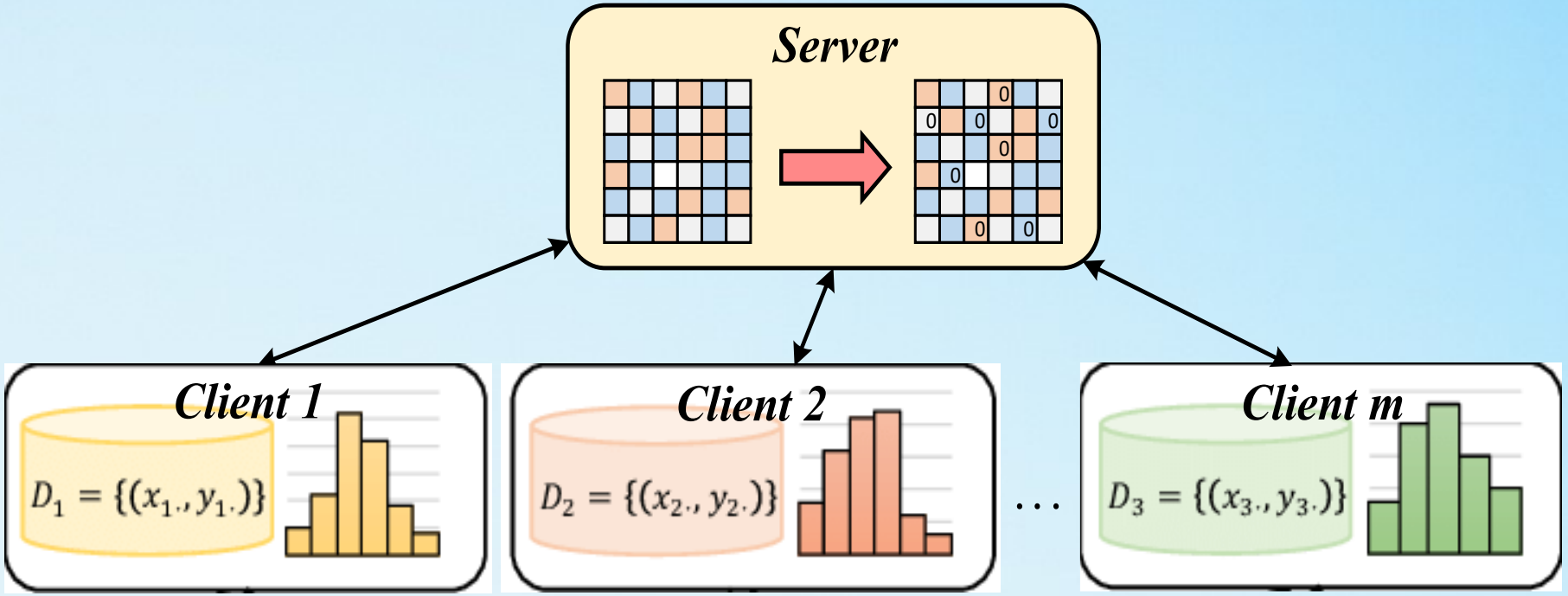
$$q_m^r = \begin{cases} q_0 \ , \ if \ r = 1 \\ \sqrt{\frac{l_{current}^r}{l_{history}^r}}\, q_m^{r-1} \ , \ otherwise \end{cases}$$

Quantize the local updates adaptively

Approach

Dynamic Global Updates Sparsification

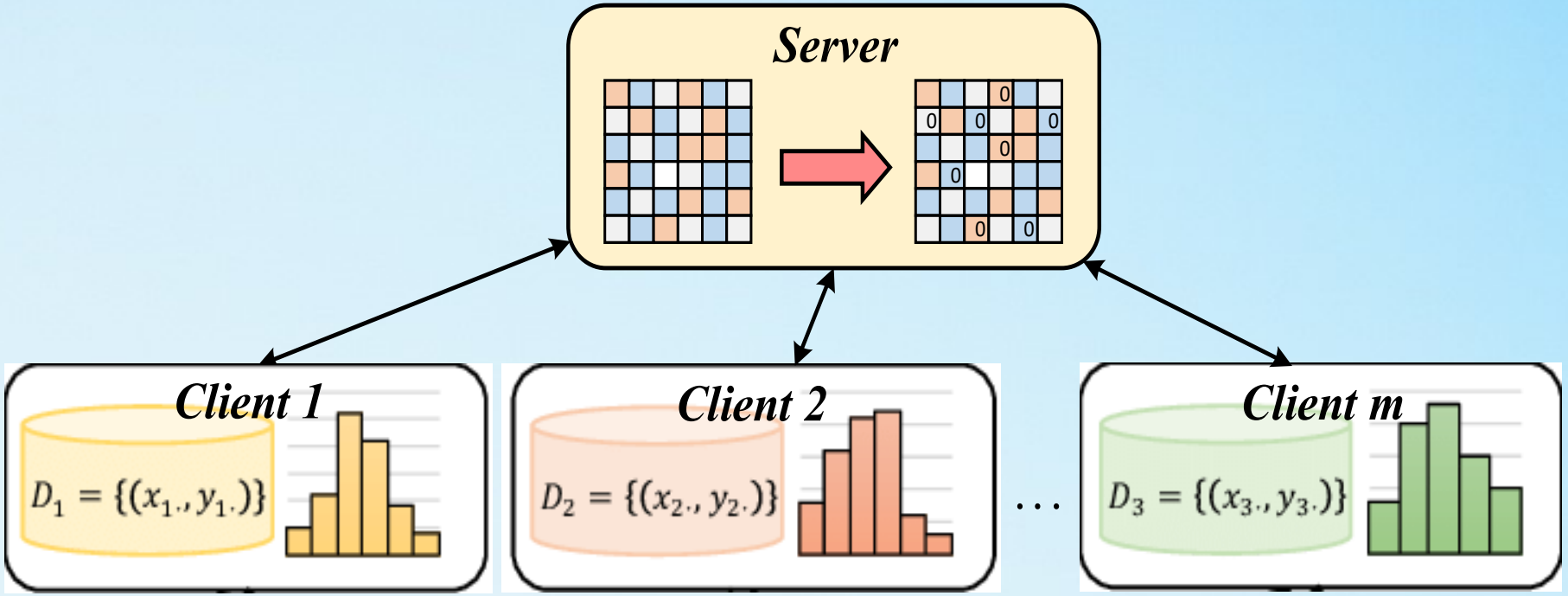
- In the initial stage of training, the data distribution among heterogeneous clients leads to a large difference between local and global updates. A smaller compression ratio should be employed to maintain the integrity of the model updates, thereby improving the training effect in the initial stage.



Approach

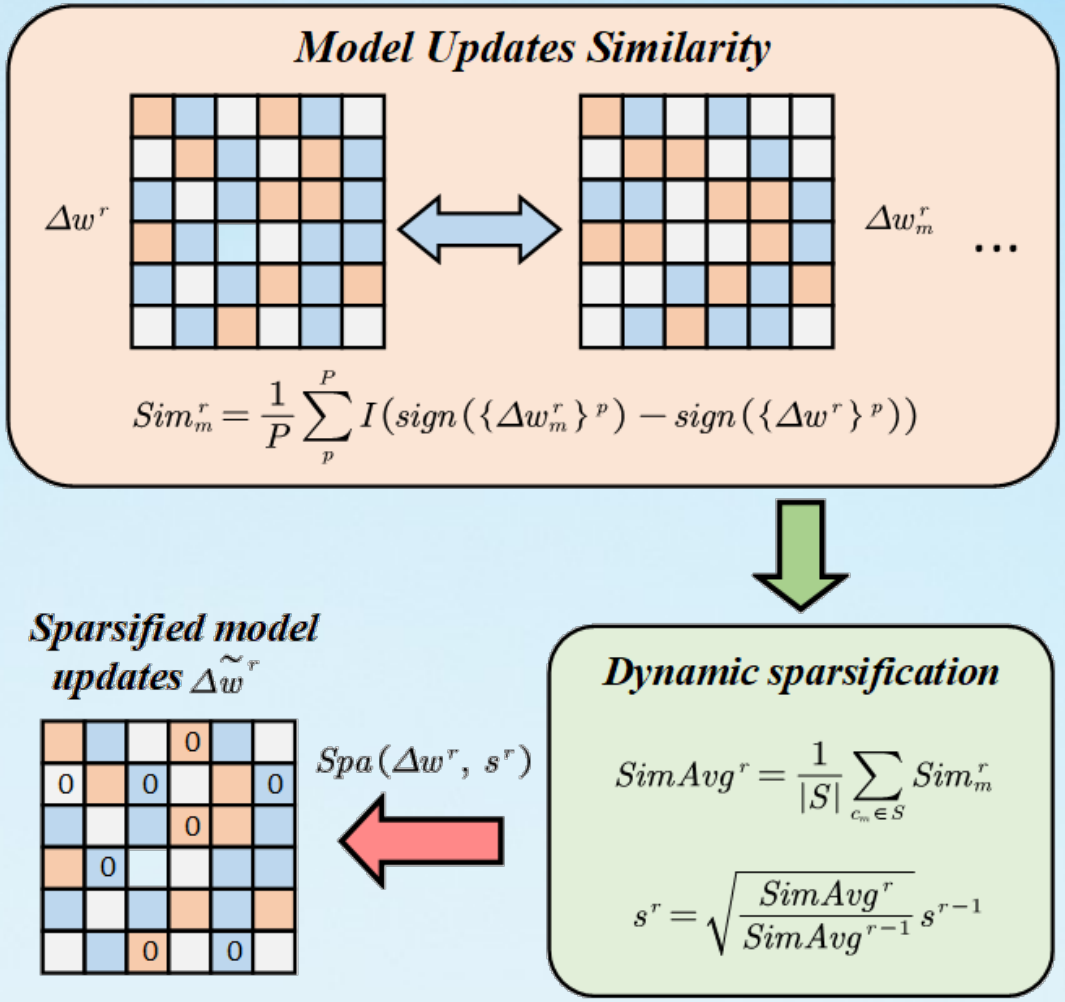
Dynamic Global Updates Sparsification

- When the global model tends to converge, the difference between local and global updates becomes smaller. The compression ratio can be scaled up to further optimize the communication efficiency.



Approach

Dynamic Global Updates Sparsification



- In the global aggregation phase, FedDAC will calculate the average of similarity.
- The number of parameters with the same update direction between local updates and global updates is utilized for similarity calculation.

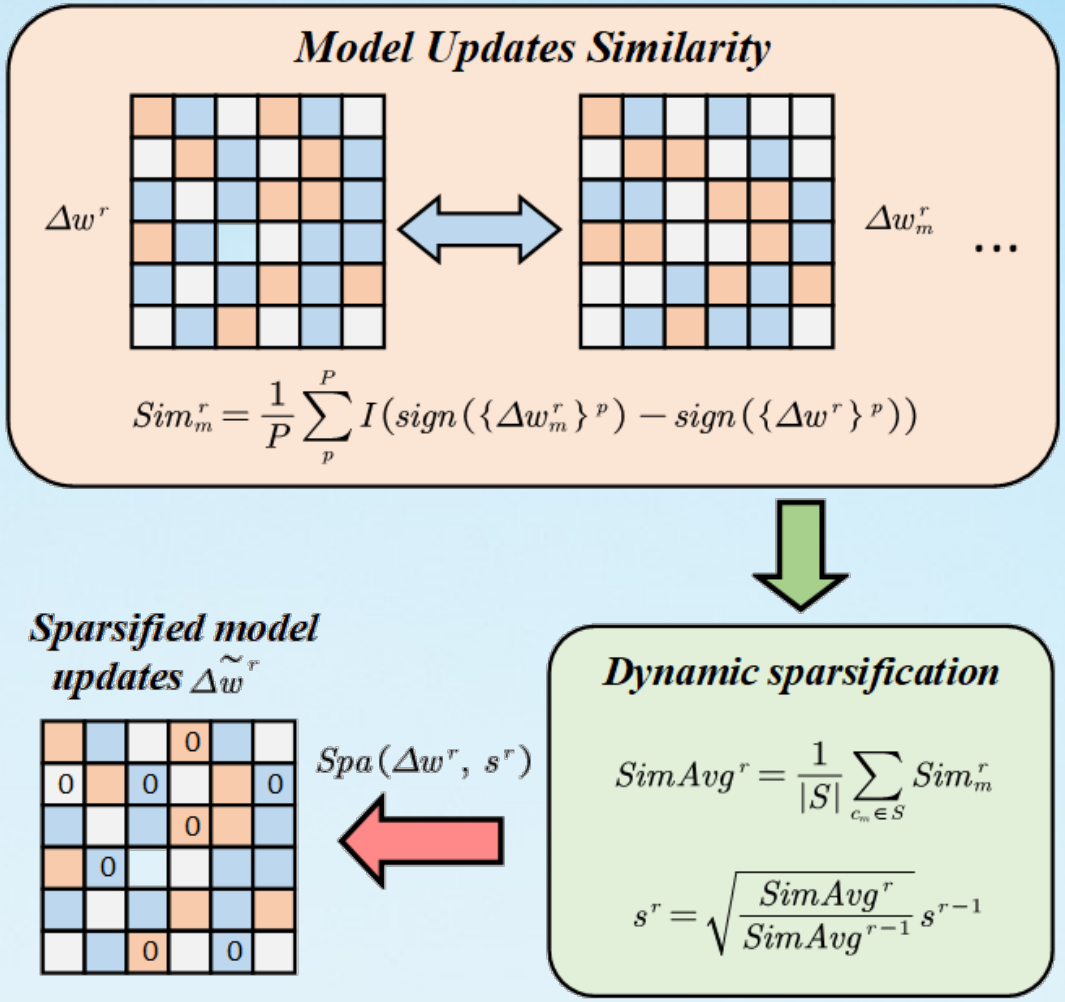
Similarity between local and global updates

$$Sim_m^r = \frac{1}{P}\sum_p^P I(sign(\{\Delta w_m^r\}^p) = sign(\{\Delta w^r\}^p))$$

$$SimAvg^r = \frac{1}{|S|}\sum_{c_m \in S} Sim_m^r$$

Approach

Dynamic Global Updates Sparsification



- The sparsity ratio is defined with the current and previous average local-global similarity, so that the server can sparsify the global model updates dynamically.

Sparsity ratio calculation

$$s^r = \begin{cases} s_0, \ if \ r = 1 \\ \sqrt{\frac{SimAvg^r}{SimAvg^{r-1}}}\, s^{r-1}, \ otherwise \end{cases}$$

Sparsify the global updates dynamically

Approach

Local and Global Residual Updates

- To alleviate the impairment of global model accuracy caused by quantization and sparsification, we introduce the residual updates on both local computation and global aggregation phases.

Client

① Compute the local residual:

$$u_m^r = \Delta w_m^r - \Delta \tilde{w}_m^r$$

② The local updates with the residual:

$$\Delta w_m^r = u_m^{r-1} - \eta \nabla f_m(w^r)$$

Server

① Compute the global residual:

$$u^r = \Delta w^r - \Delta \tilde{w}^r$$

② The global updates with the residual:

$$\Delta w^r = u^{r-1} + \sum_{c_m \in S} \frac{1}{|S|} \triangle \tilde{w}_m^r$$

Approach

Local and Global Residual Updates

- The residual refers to the difference between the full precision model and the lossy compressed model. The local residual is introduced into the local updates and the global updates is composed of its original values and the global residual.

Client

① Compute the local residual:

$$u_m^r = \Delta w_m^r - \Delta \tilde{w}_m^r$$

② The local updates with the residual:

$$\Delta w_m^r = u_m^{r-1} - \eta \nabla f_m(w^r)$$

Server

① Compute the global residual:

$$u^r = \Delta w^r - \Delta \tilde{w}^r$$

② The global updates with the residual:

$$\Delta w^r = u^{r-1} + \sum_{c_m \in S} \frac{1}{|S|} \triangle \tilde{w}_m^r$$

Experiment

Experiment Setup

- We conduct experiments to evaluate the performance of FedDAC.
- Specifically, we choose the MNIST and CIFAR-10 as experiment datasets.
- We employ the Dirichlet distribution simulate the data heterogeneity.

Dataset
MNIST
CIFAR-10

Experiment Setup

- The parameter settings are shown in the table.

Parameter	Value
The iteration rounds: R	200
The local learning rate: η_l	0.1/0.01
The number of clients: M	100
Clients selected for in each round	10
The initial quantization coefficient: q_0	64/128
The initial sparsity ratio s_0	0.2/0.1

- And the indicators include **the accumulated communication volume** and **the global model accuracy**.

Experiment

■ Analysis of the hyperparameter selection

- To choose suitable hyperparameter μ for the subsequent experiments, we analysis the hyperparameter under different situations. The experiments results are shown below.

TABLE III
ACCUMULATED COMMUNICATION VOLUME (MB) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	12.0	11.3	**1.4**	**215.0**	**163.8**	**153.6**
20	**10.8**	**8.9**	1.5	225.3	184.3	163.8
30	16.9	16.1	3.2	245.8	225.3	204.8
40	32.3	29.0	5.7	358.4	317.4	256.0
50	63.2	59.6	11.9	409.6	389.1	378.9

TABLE IV
GLOBAL MODEL ACCURACY (%) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	86.84	90.09	90.37	52.86	60.19	**63.33**
20	86.87	90.10	90.36	52.90	60.22	63.30
30	86.86	90.12	90.35	52.84	60.21	63.28
40	86.86	90.13	90.38	52.87	60.23	63.32
50	**86.88**	**90.15**	**90.41**	**52.91**	**60.26**	63.31

Experiment

■ Analysis of the hyperparameter selection

- The left table shows the accumulated communication volume with different μ. We can see when μ equals to 10, it has the smallest accumulated communication volume under all situations. As μ getting larger, the accumulated communication volume increases significantly.

TABLE III
ACCUMULATED COMMUNICATION VOLUME (MB) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	12.0	11.3	**1.4**	**215.0**	**163.8**	**153.6**
20	**10.8**	**8.9**	1.5	225.3	184.3	163.8
30	16.9	16.1	3.2	245.8	225.3	204.8
40	32.3	29.0	5.7	358.4	317.4	256.0
50	63.2	59.6	11.9	409.6	389.1	378.9

TABLE IV
GLOBAL MODEL ACCURACY (%) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	86.84	90.09	90.37	52.86	60.19	**63.33**
20	86.87	90.10	90.36	52.90	60.22	63.30
30	86.86	90.12	90.35	52.84	60.21	63.28
40	86.86	90.13	90.38	52.87	60.23	63.32
50	**86.88**	**90.15**	**90.41**	**52.91**	**60.26**	63.31

Experiment

■ Analysis of the hyperparameter selection

- The right table shows the global model accuracy with different μ. Under week data heterogeneous situation that is α equals to 10, it has the highest global model accuracy when μ equals 10.

TABLE III
ACCUMULATED COMMUNICATION VOLUME (MB) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	12.0	11.3	**1.4**	**215.0**	**163.8**	**153.6**
20	**10.8**	**8.9**	1.5	225.3	184.3	163.8
30	16.9	16.1	3.2	245.8	225.3	204.8
40	32.3	29.0	5.7	358.4	317.4	256.0
50	63.2	59.6	11.9	409.6	389.1	378.9

TABLE IV
GLOBAL MODEL ACCURACY (%) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	86.84	90.09	90.37	52.86	60.19	**63.33**
20	86.87	90.10	90.36	52.90	60.22	63.30
30	86.86	90.12	90.35	52.84	60.21	63.28
40	86.86	90.13	90.38	52.87	60.23	63.32
50	**86.88**	**90.15**	**90.41**	**52.91**	**60.26**	63.31

Experiment

■ Analysis of the hyperparameter selection

- Based on the above analysis, we set μ to 10, Therefore, the highest communication efficiency can be obtained with a slight decrease in global model accuracy.

TABLE III
ACCUMULATED COMMUNICATION VOLUME (MB) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	12.0	11.3	**1.4**	**215.0**	**163.8**	**153.6**
20	**10.8**	**8.9**	1.5	225.3	184.3	163.8
30	16.9	16.1	3.2	245.8	225.3	204.8
40	32.3	29.0	5.7	358.4	317.4	256.0
50	63.2	59.6	11.9	409.6	389.1	378.9

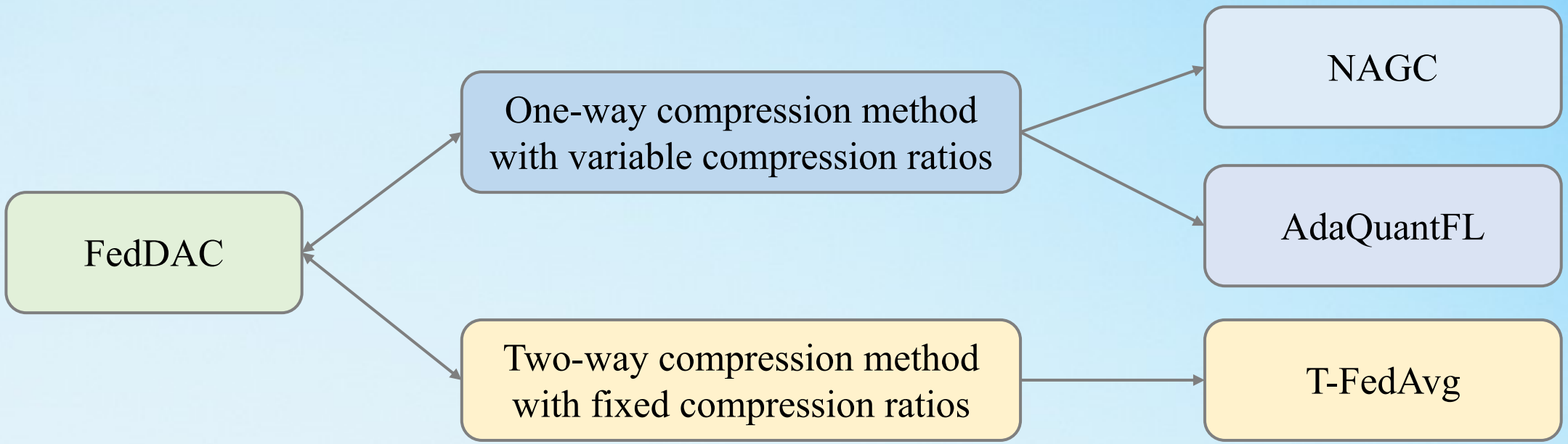
TABLE IV
GLOBAL MODEL ACCURACY (%) WITH μ

μ	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
10	86.84	90.09	90.37	52.86	60.19	**63.33**
20	86.87	90.10	90.36	52.90	60.22	63.30
30	86.86	90.12	90.35	52.84	60.21	63.28
40	86.86	90.13	90.38	52.87	60.23	63.32
50	**86.88**	**90.15**	**90.41**	**52.91**	**60.26**	63.31

Experiment

■ Analysis of the accumulated communication volume

- After the hyperparameter selection, we compare FedDAC with NAGC、AdaQuantFL, which are one-way compression methods considering variable compression ratios and T-FedAvg, which is a two-way compression method with a fixed compression coefficient.



Experiment

■ Analysis of the accumulated communication volume

- This figure shows the accumulated communication volume on MNIST of the four methods.
- We can see that FedDAC can achieve the smallest accumulated communication volume.

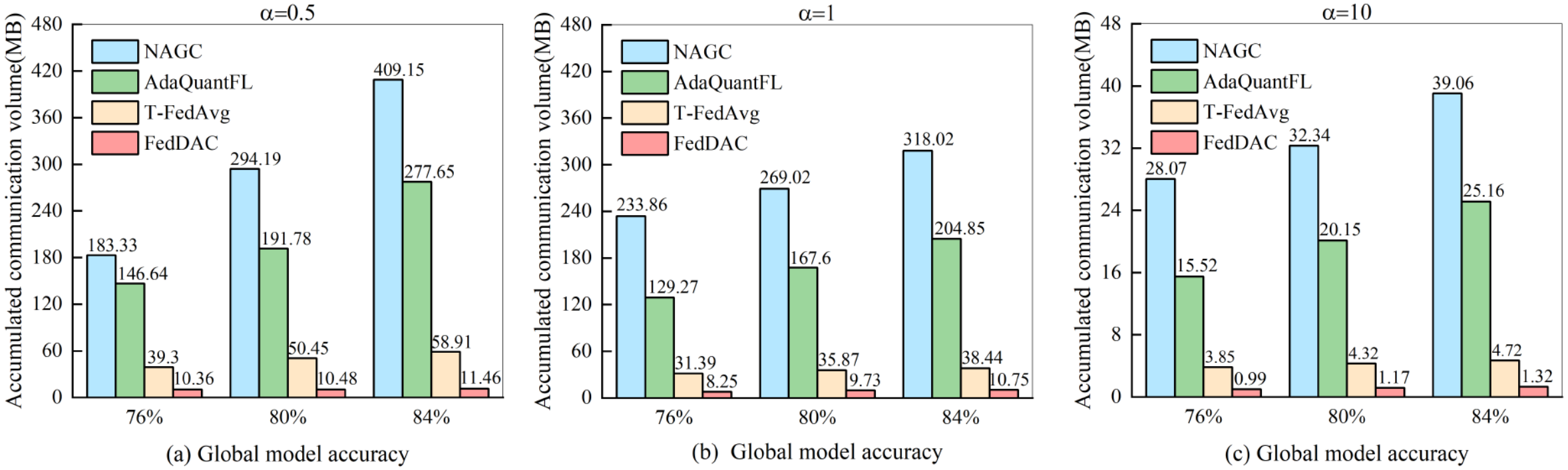


Fig. 5. Accumulated communication volume on the heterogeneous MNIST dataset for NAGC, AdaQuantFL, T-FedAvg, and FedDAC

■ Analysis of the accumulated communication volume

- The next figure shows the accumulated communication volume on CIFAR-10 of the four methods. Same as the results on MNIST, our method can still achieve the smallest accumulated communication volume.

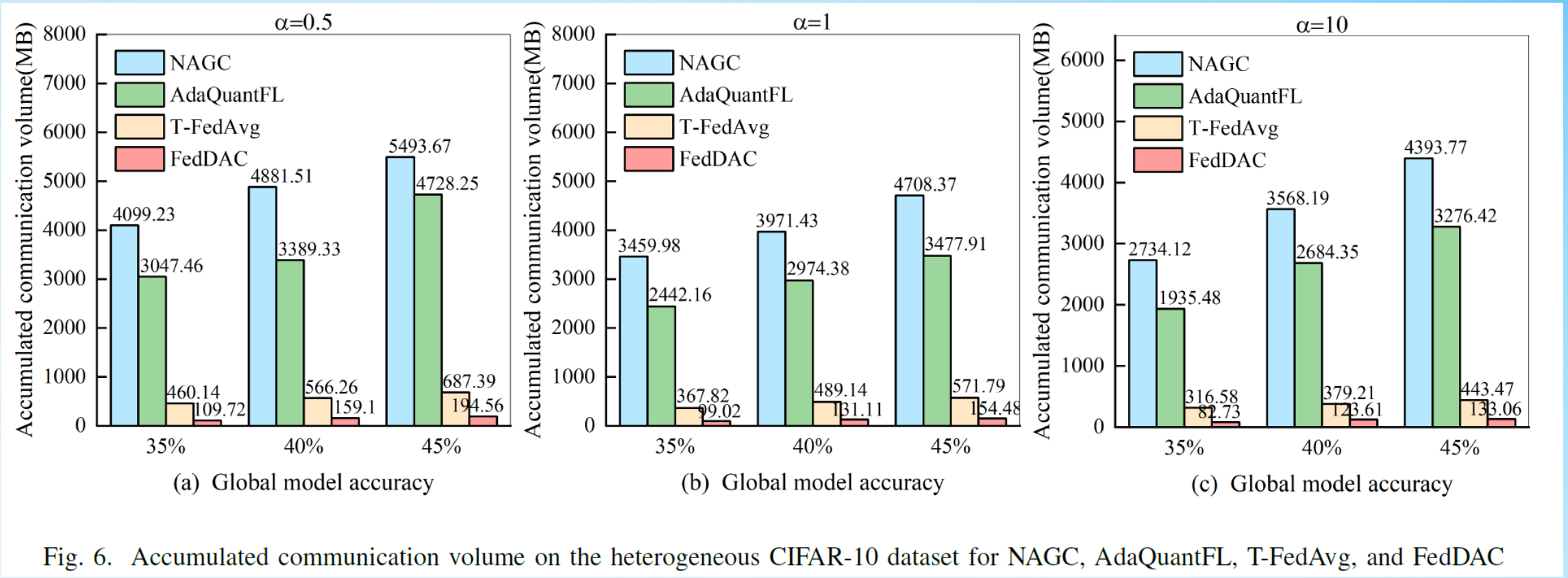


Fig. 6. Accumulated communication volume on the heterogeneous CIFAR-10 dataset for NAGC, AdaQuantFL, T-FedAvg, and FedDAC

Experiment

■ Analysis of the global model accuracy

- We also compare the global model accuracy of FedDAC with other three methods.
- The results are shown in this table.

TABLE V
GLOBAL MODEL ACCURACY OF DIFFERENT APPROACHES

Comparison approaches	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
NAGC	84.20	87.78	88.82	49.41	57.39	61.25
AdaQuantFL	86.34	89.42	**91.55**	**53.02**	**61.47**	62.32
T-FedAvg	84.29	88.17	89.37	46.76	54.29	60.88
FedDAC	**86.84**	**90.09**	90.37	51.86	60.19	**63.33**

Experiment

■ Analysis of the global model accuracy

- On the MNIST dataset, FedDAC outperforms the other three methods under strong data heterogeneity situation, that is α equals to 0.5 or 1.

TABLE V
GLOBAL MODEL ACCURACY OF DIFFERENT APPROACHES

Comparison approaches	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
NAGC	84.20	87.78	88.82	49.41	57.39	61.25
AdaQuantFL	86.34	89.42	**91.55**	**53.02**	**61.47**	62.32
T-FedAvg	84.29	88.17	89.37	46.76	54.29	60.88
FedDAC	**86.84**	**90.09**	90.37	51.86	60.19	**63.33**

Experiment

■ Analysis of the global model accuracy

- On the CIFAR-10 dataset, under week data heterogeneity situation, that is α equals to 10, FedDAC outperforms the other three methods.

TABLE V
GLOBAL MODEL ACCURACY OF DIFFERENT APPROACHES

Comparison approaches	Dataset					
	MNIST			CIFAR-10		
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 10$
NAGC	84.20	87.78	88.82	49.41	57.39	61.25
AdaQuantFL	86.34	89.42	**91.55**	**53.02**	**61.47**	62.32
T-FedAvg	84.29	88.17	89.37	46.76	54.29	60.88
FedDAC	**86.84**	**90.09**	90.37	51.86	60.19	**63.33**

Conclusion

Conclusion

- In order to reduce the significant communication costs in heterogeneous federated learning while achieving the trade-off between communication efficiency and global model accuracy, a ***Dual Adaptive Compression*** method (***FedDAC***) is proposed in this paper.
  - In the local computation phase, the loss queue is adopted to detect the convergence trends within each client. FedDAC can then dynamically quantify model updates and allow for various compression ratios among heterogeneous clients.
  - In the global aggregation phase, FedDAC can determine the fluctuations in training based on the similarity between clients and the server, thereby adjusting the sparsity ratio flexibly.
  - To alleviate the reduction in model accuracy caused by lossy compression, we introduce residual updates in the local computation and global aggregation phases to maintain model accuracy.

Future Work

- Experiments on large-scale datasets
- Further improve the accuracy

At Last

Thanks for your attention!