INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING
ICPP / 2021 / CHICAGO / USA
acm In-Cooperation
sighpc
AUGUST 9-12, 2021

Dubhe: Towards Data Unbiasedness with Homomorphic Encryption in Federated Learning Client Selection

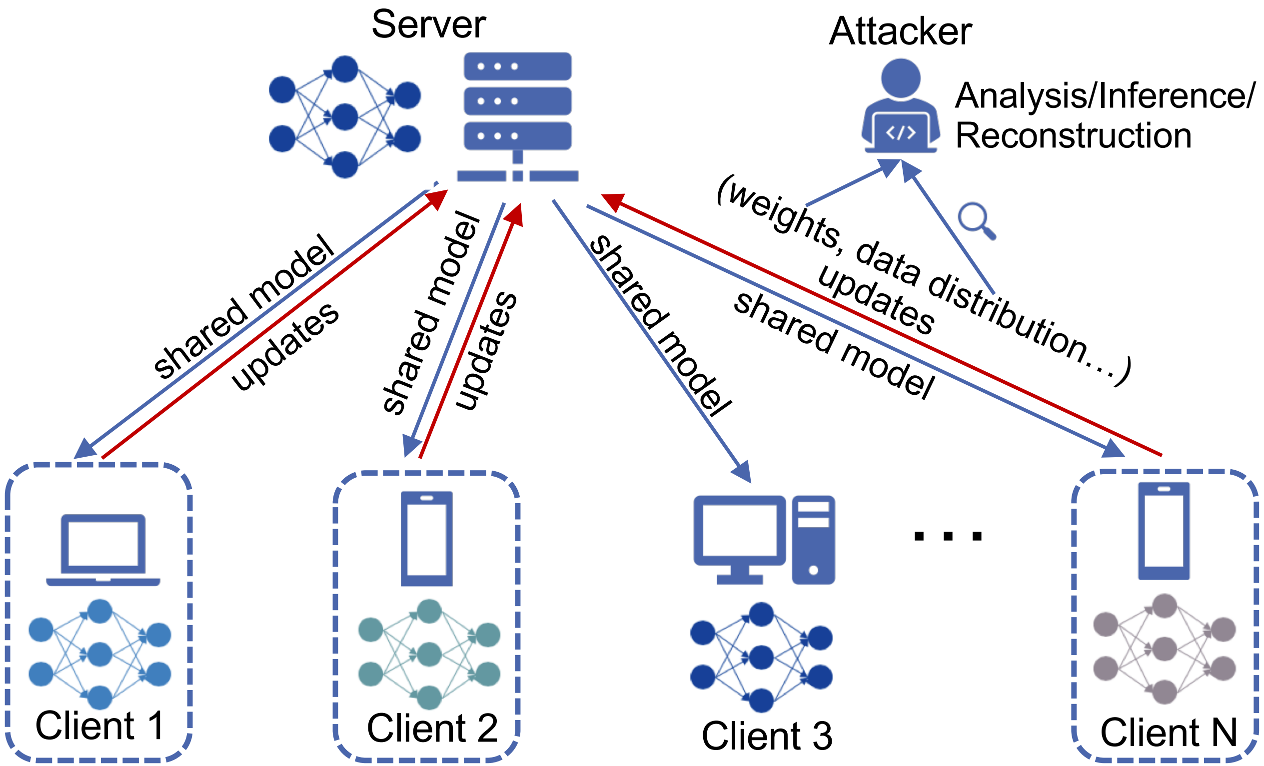
Shulai Zhang, Zirui li, Quan Chen, Wenli Zheng, Jingwen Leng, Minyi Guo

Shanghai Jiao Tong University

SHANGHAI JIAO TONG UNIVERSITY

Federated Learning

- ***Demand:** Security promise*
  - Homomorphic Encryption
- ***Problem:** Statistical heterogeneity*
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness
- ***Method:** Client selection*
  - Random selection
  - Greedy selection

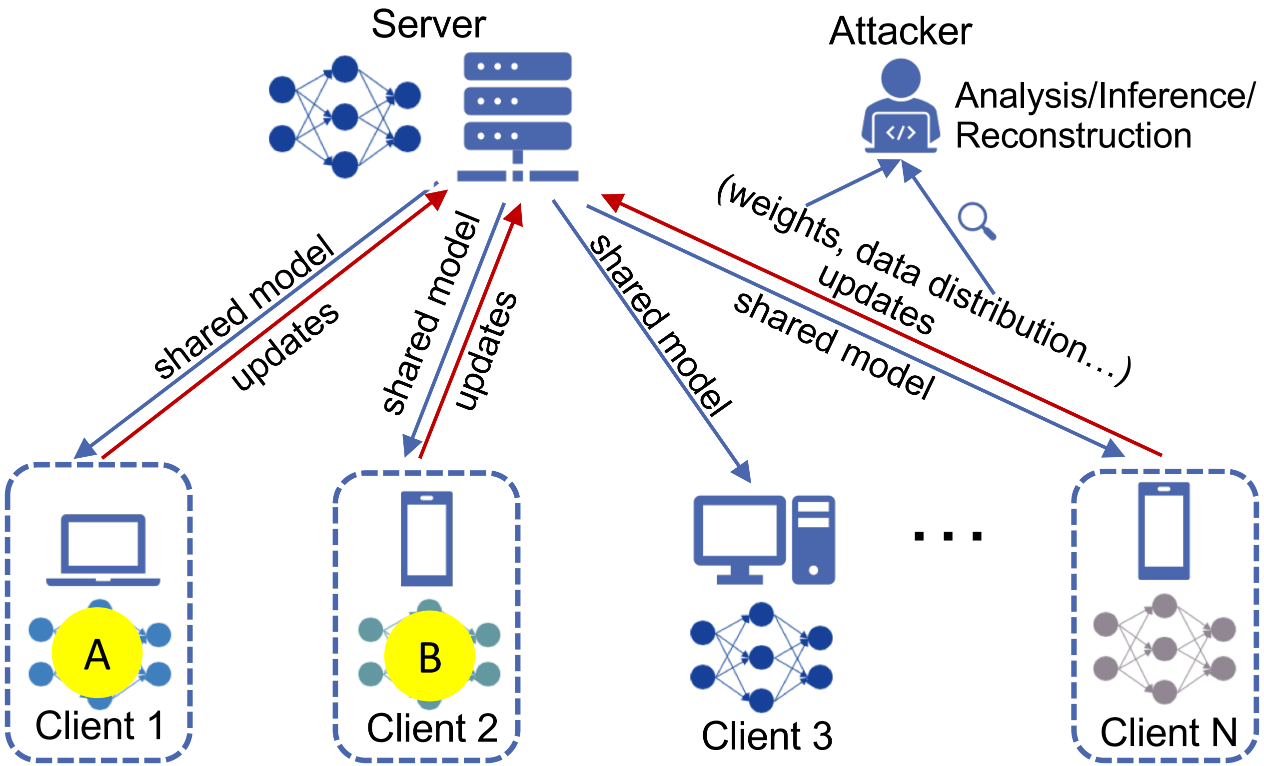


The framework of Federated Learning

Background

Federated Learning

- ***Demand:*** *Security promise*
  - **Homomorphic Encryption**
- ***Problem:*** *Statistical heterogeneity*
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness
- ***Method:*** *Client selection*
  - Random selection
  - Greedy selection

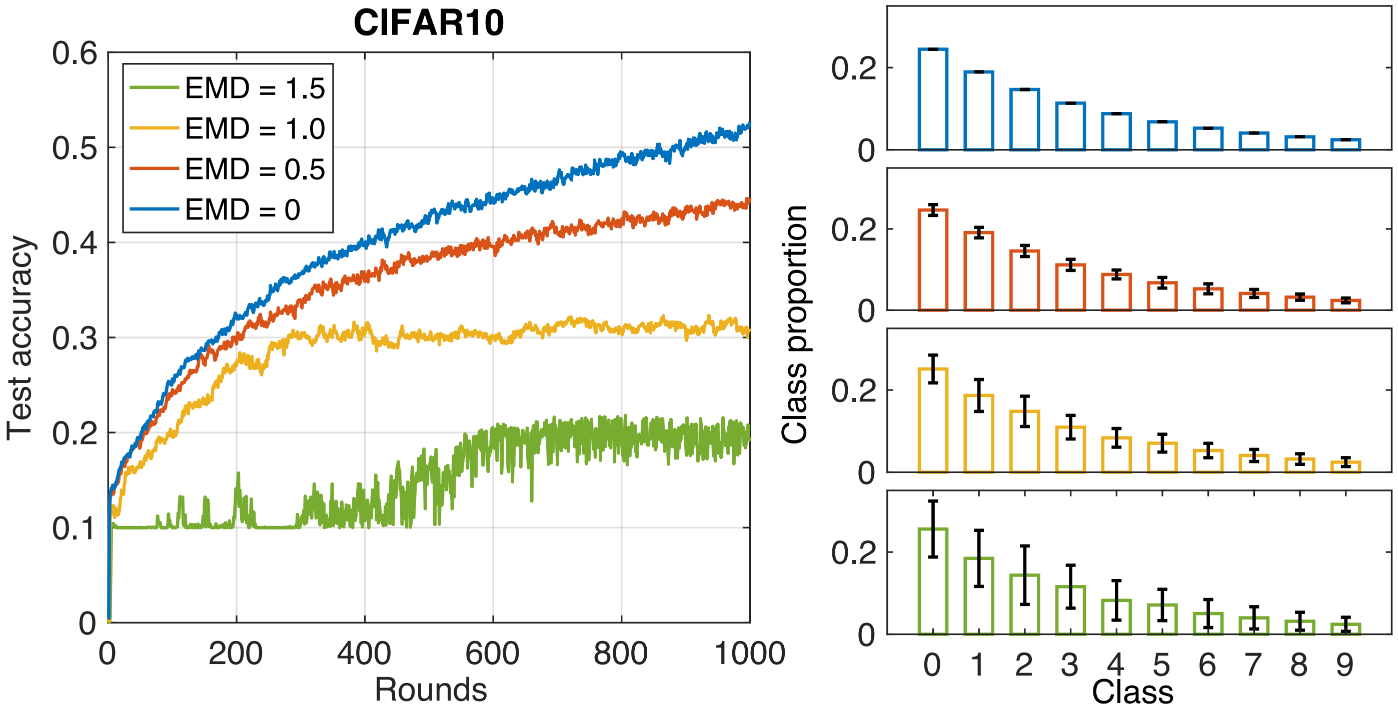


The framework of Federated Learning

Motivation

Federated Learning

- ***Demand:*** *Security promise*
  - Homomorphic Encryption
- ***Problem:*** *Statistical heterogeneity*
  - Local data skewness
  - **Data discrepancy among clients**
  - Global data skewness
- ***Method:*** *Client selection*
  - Random selection
  - Greedy selection



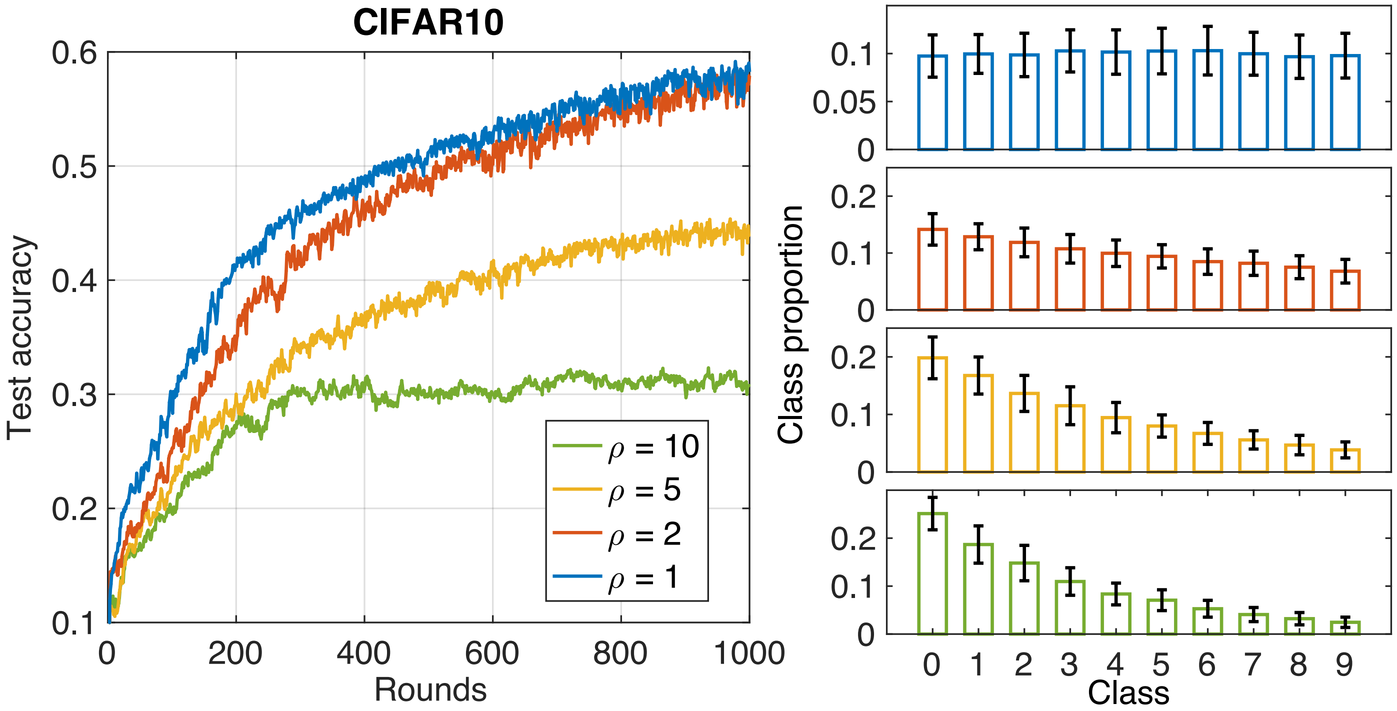
Discrepancy of clients in CIFAR10 classification

SHANGHAI JIAO TONG UNIVERSITY

Motivation

Federated Learning

- ***Demand:** Security promise*
  - Homomorphic Encryption
- ***Problem:** Statistical heterogeneity*
  - Local data skewness
  - Data discrepancy among clients
  - **Global data skewness**
- ***Method:** Client selection*
  - Random selection
  - Greedy selection



Global data skewness in CIFAR10 classification

Mathematical Demonstration

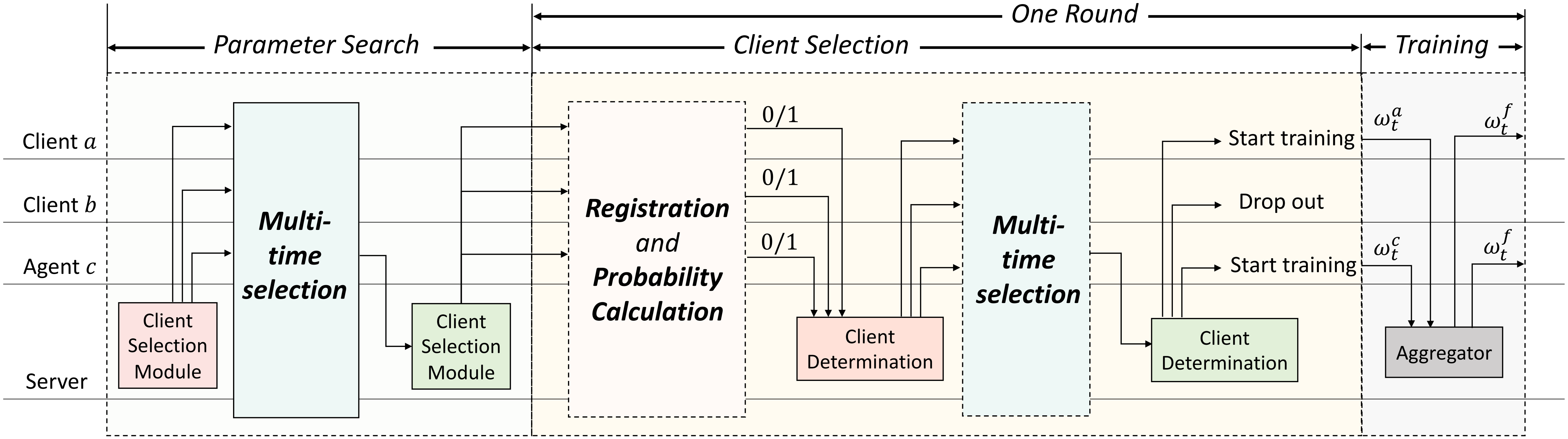
Federated Learning

- ***Demand:*** *Security promise*
  - Homomorphic Encryption
- ***Problem:*** *Statistical heterogeneity*
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness
- ***Method:*** *Client selection*
  - Random selection
  - Greedy selection

$$
\begin{aligned}
||\omega^f_{mT} - \omega^*_{mT}|| \leq & ||\omega^f_{mT} - \omega^c_{mT} + \omega^c_{mT} - \omega^*_{mT}|| \\
\leq & \frac{1}{K}\sum_{i=1}^{K}[(1+\eta\lambda)^T ||\omega^k_{(m-1)T} - \omega^c_{(m-1)T}|| \\
& + \underbrace{\eta||p^k_l - p_o||_1}_{①}(\sum_{j=2}^{T} \mathrm{g}(\omega^c_{mT-j})(1+\eta\lambda)^{j-1})] \\
& + (1+\eta\lambda)^T ||\omega^c_{(m-1)T} - \omega^*_{(m-1)T}|| \\
& + \underbrace{\eta||p_o - p_u||_1}_{②}(\sum_{j=1}^{T} \mathrm{g}(\omega^*_{mT-j})(1+\eta\lambda)^{j-1}),
\end{aligned}
$$

Target: $\min||p_o - p_u||_1$

The Design of Dubhe



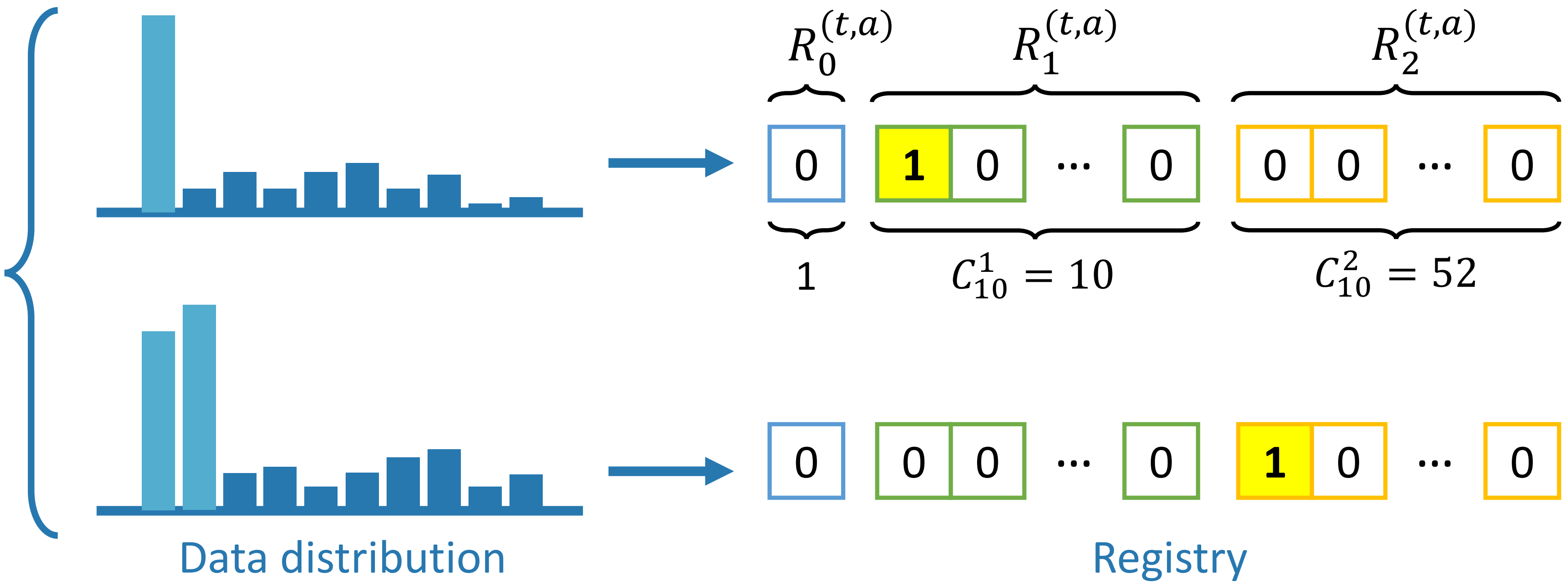
The framework of Dubhe

The Design of Dubhe

Registration *and* ***Probability Calculation***

- The ***registry*** encodes the client's data distribution information in a one-hot manner.
- In classification problems, the registry encodes each client's data distribution by its ***dominating classes***.

Samples of a 10-class classification



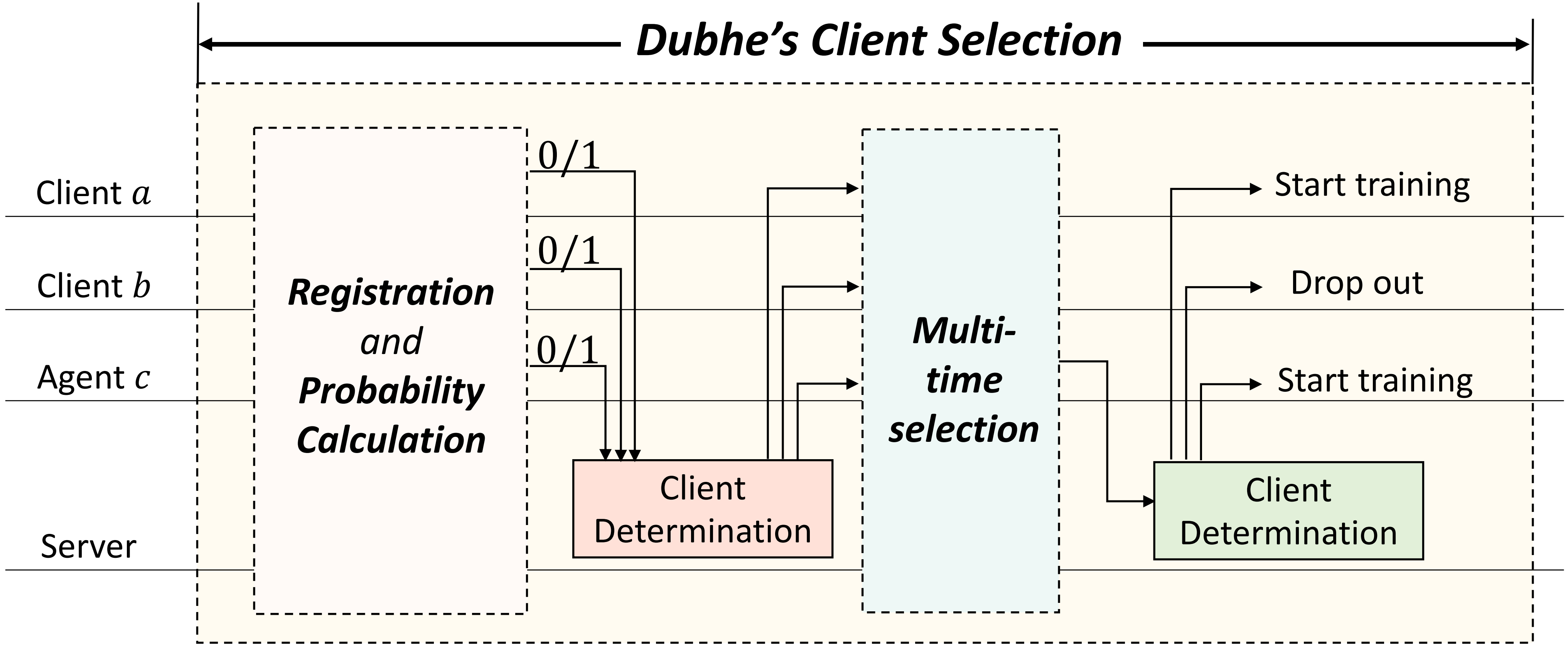
INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING

50th International Conference on Parallel Processing (ICPP)
August 9-12, 2021 in Virtual Chicago, IL

acm In-Cooperation sighpc

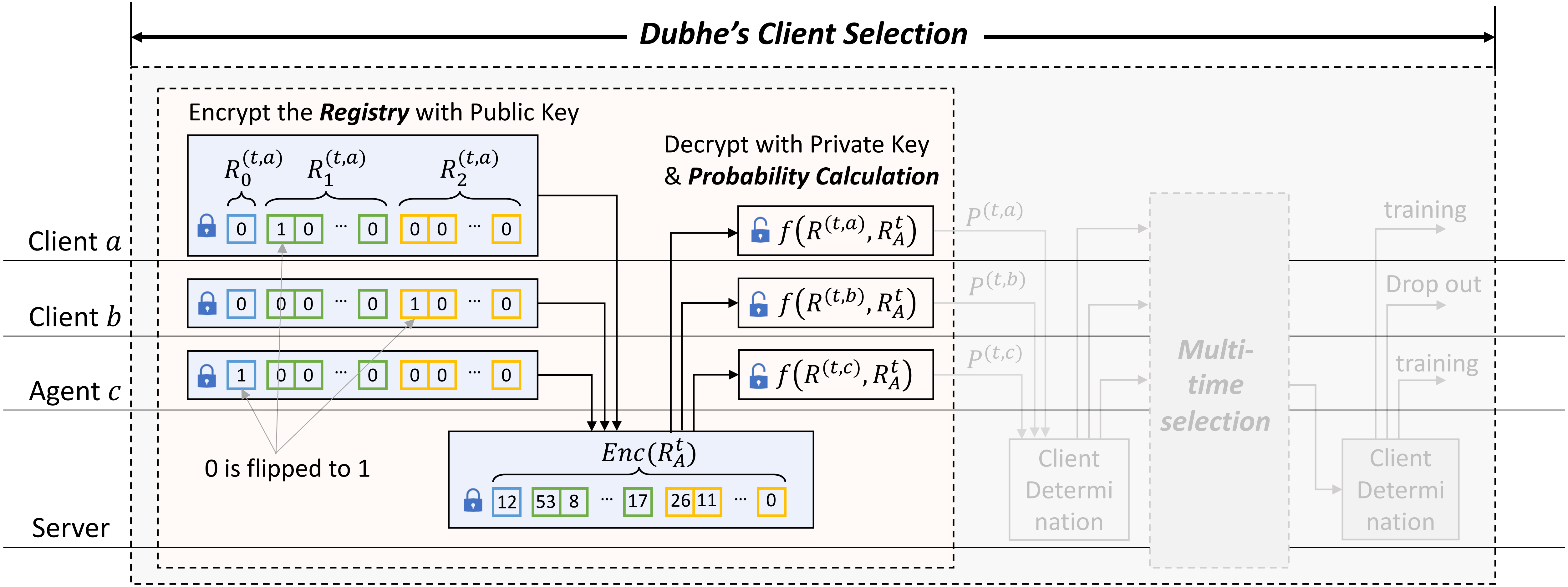
The Design of Dubhe

Registration and *Probability Calculation*

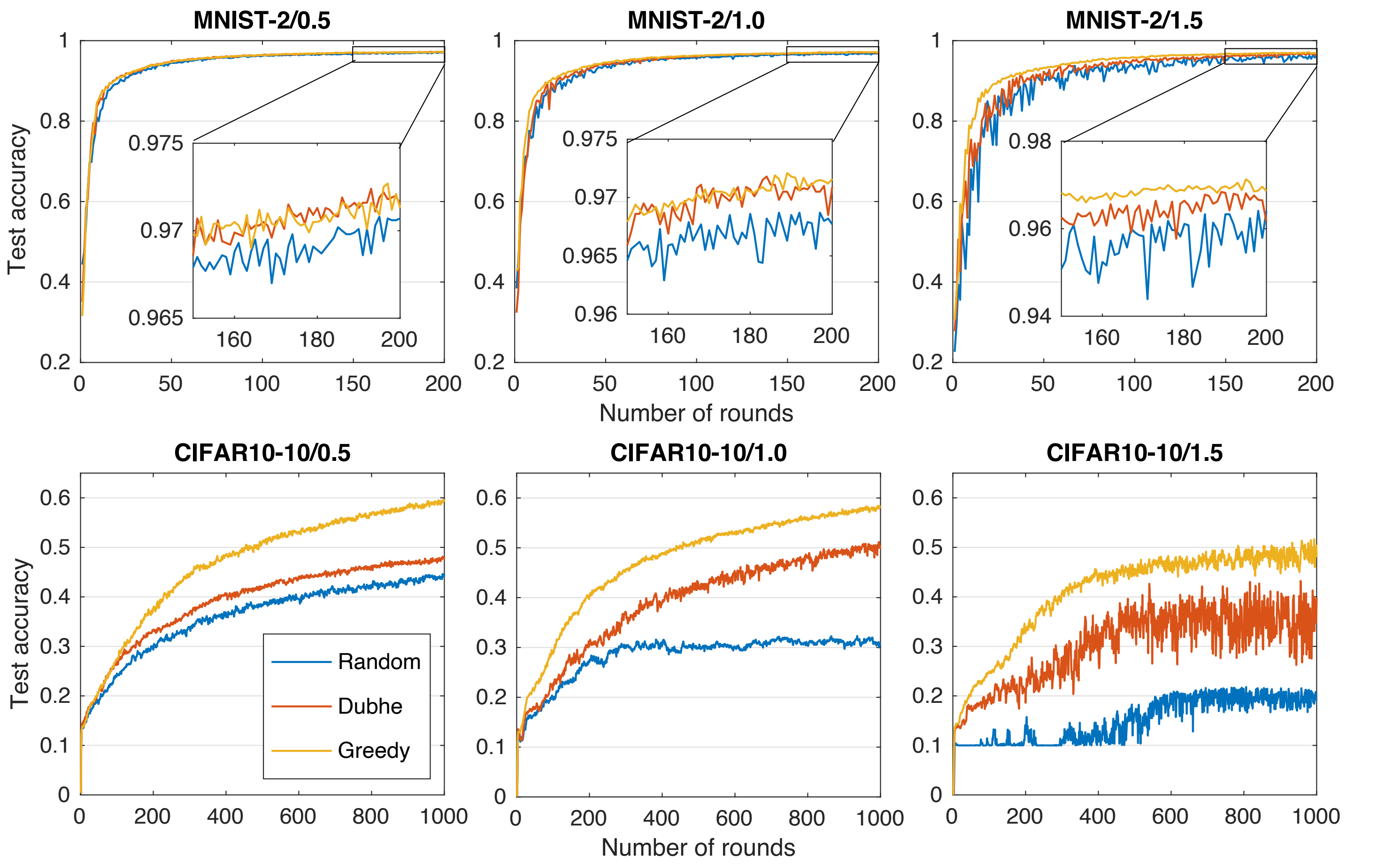


The Design of Dubhe

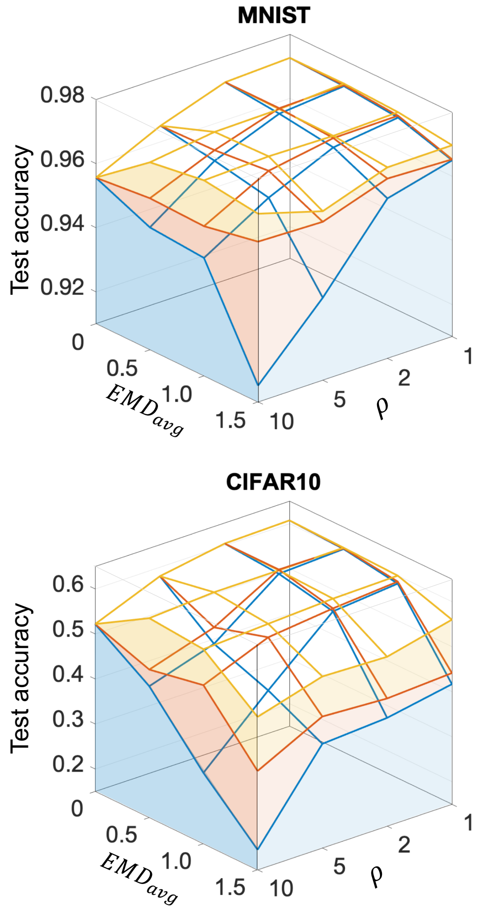
Registration** and **Probability Calculation



Evaluation of Dubhe

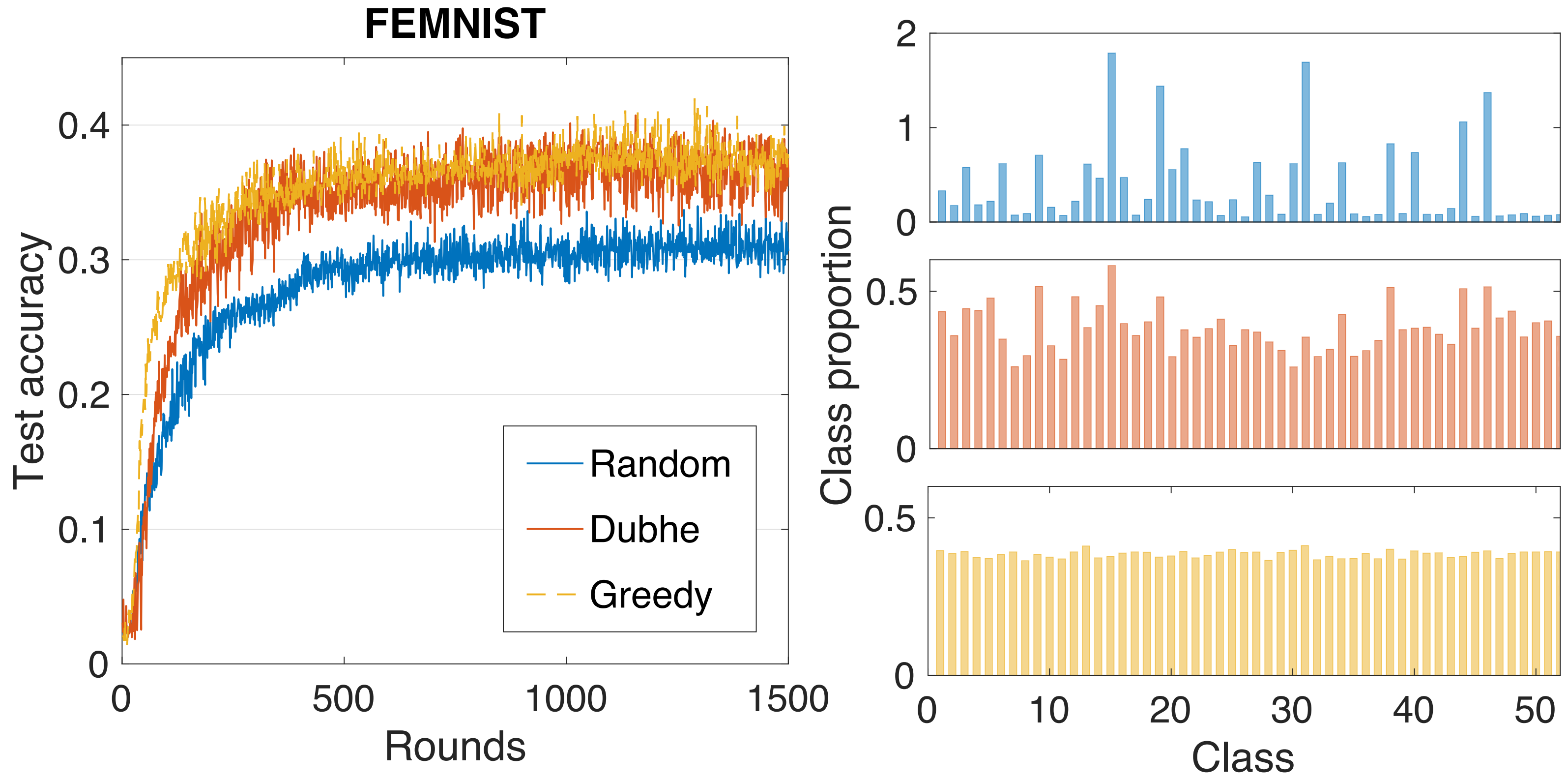


Test accuracy curves on MNIST and CIFAR10



Average accuracy

Evaluation of Dubhe

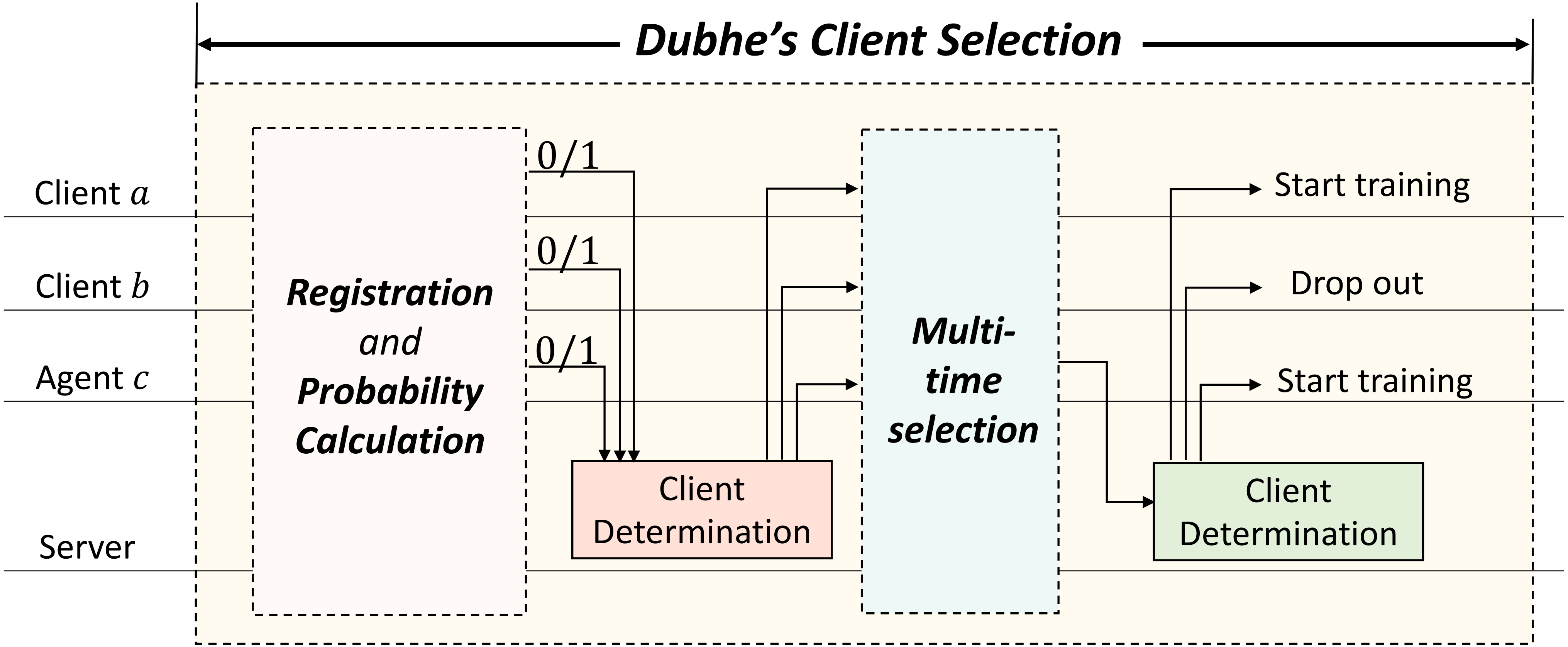


Results on FEMNIST

- The data balancing performance of Dubhe is approaching the performance of the greedy selection.

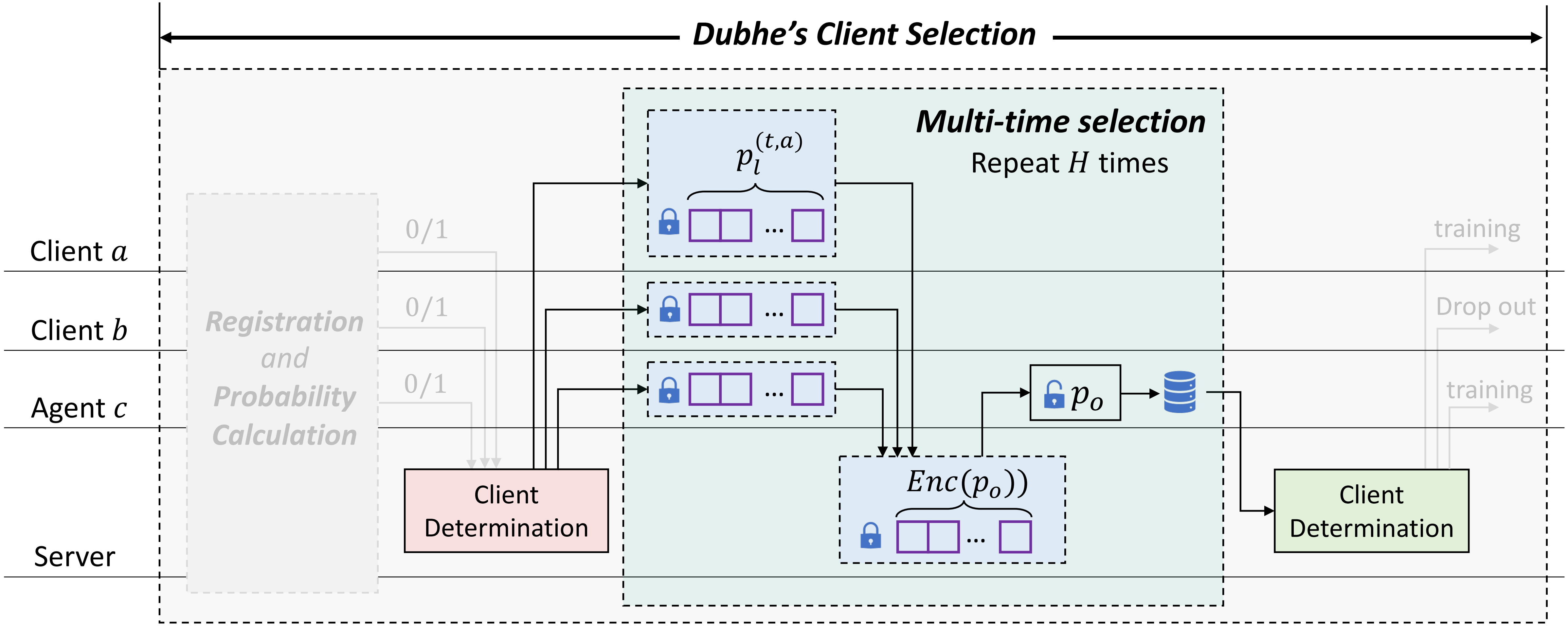
The Design of Dubhe

Multi-time selection



The Design of Dubhe

Multi-time selection



Evaluation of Dubhe

Multi-time selection

SHANGHAI JIAO TONG UNIVERSITY

Results with multi-time selection. *M* refers to MNIST and *C* refers to CIFAR10.

H	1	2	5	10	20	opt
EMD^*	0.2946	0.2588	0.2176	0.1971	**0.1750**	0.0144
Acc^M	0.9662	0.9668	0.9665	**0.9684**	0.9678	0.9694
β^M	0.0%	17.6%	10.5%	**69.5%**	51.5%	100%
Acc^C	0.4300	0.4518	0.4486	0.4441	**0.4577**	0.5295
β^C	0.00%	14.8%	12.6%	9.5%	**18.8%**	100%

EMD^* decreases with larger H, thereby improving the model accuracy

Evaluation of Dubhe

SHANGHAI JIAO TONG UNIVERSITY

Registry sparsity

$R_{A,0}$

(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
179	130	89	60	39	18	8	2	1	0

$R_{A,1}$

(0,1) 22	(0,2) 20	(0,3) 19	(0,4) 19	(0,5) 21	(0,6) 13	(0,7) 13	(0,8) 14	(0,9) 9
(1,2) 13	(1,3) 20	(1,4) 16	(1,5) 14	(1,6) 16	(1,7) 13	(1,8) 10	(1,9) 8	
(2,3) 20	(2,4) 19	(2,5) 16	(2,6) 12	(2,7) 11	(2,8) 8	(2,9) 3		
(3,4) 12	(3,5) 9	(3,6) 12	(3,7) 9	(3,8) 5	(3,9) 8			
(4,5) 7	(4,6) 6	(4,7) 6	(4,8) 7	(4,9) 4				
(5,6) 4	(5,7) 9	(5,8) 6	(5,9) 2					
(6,7) 6	(6,8) 3	(6,9) 7						
(7,8) 2	(7,9) 1							
(8,9) 0								

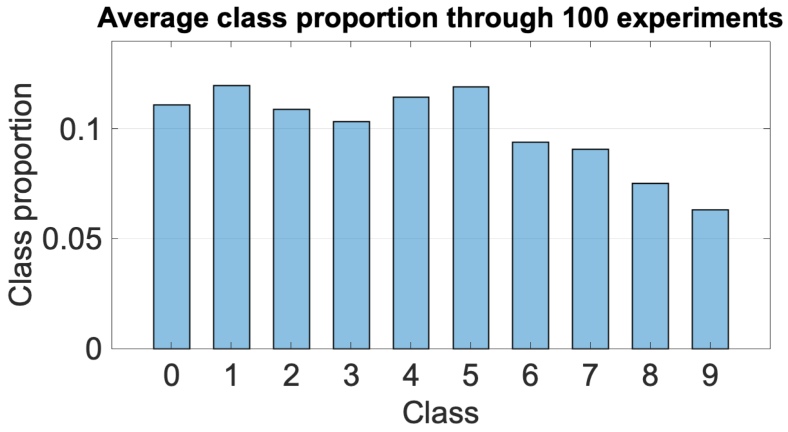
$R_{A,10}$

None
0

There are 3 out of 1000 clients who have class 2 and class 9 dominated.

$$\kappa\big(u = (2,9)\big) = 3$$

- An example of the overall registry and its corresponding participated class proportion.



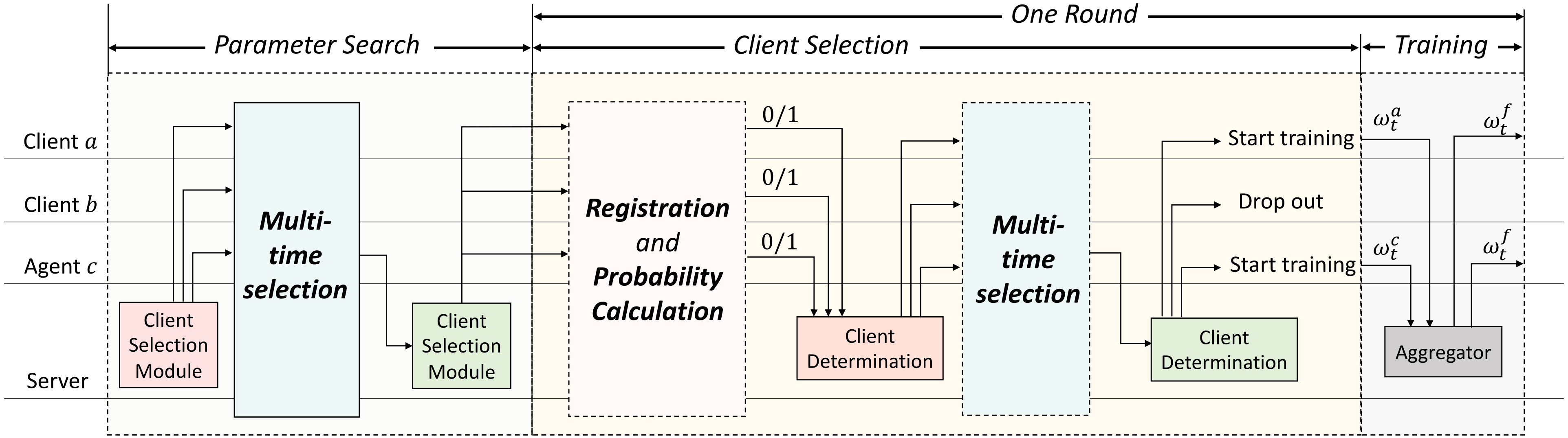
The sparsity of the overall registry is the primary cause of this deviation from expectation.

INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING

acm In-Cooperation sighpc

Conclusions

- The impact of data skewness on the performance degradation in FL is mathematically demonstrated.
- We propose ***Dubhe***, a proactive client selection system to balance skewed data
  - pluggable, adaptive and robust to various FL settings
  - with negligible encryption and communication overhead
  - improves the training performance without bringing security threats.



Thanks for listening!

Any questions? Comments? Please contact:

Shulai Zhang: zslzsl1998@sjtu.edu.cn