FedGreen: Federated Learning with Fine-Grained Gradient Compression for Green Mobile Edge Computing


Peichun Li[1]


Xumin Huang[1]


Miao Pan[2]


Rong Yu[1]

[1]*Guangdong University of Technology* [2]*University of Houston*

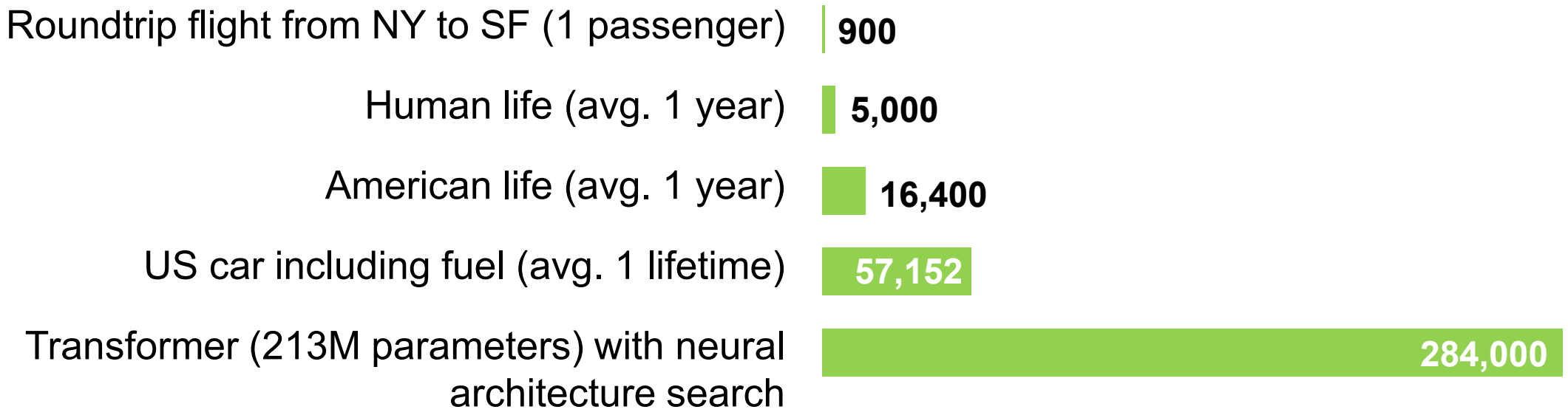
IEEE GLOBECOM 2021

Project page http://fedgreen.find-lab.com/

Background: The Demand for Efficient FL in MEC

find

- **Common carbon footprint benchmarks (in KG of CO2 equivalent)**

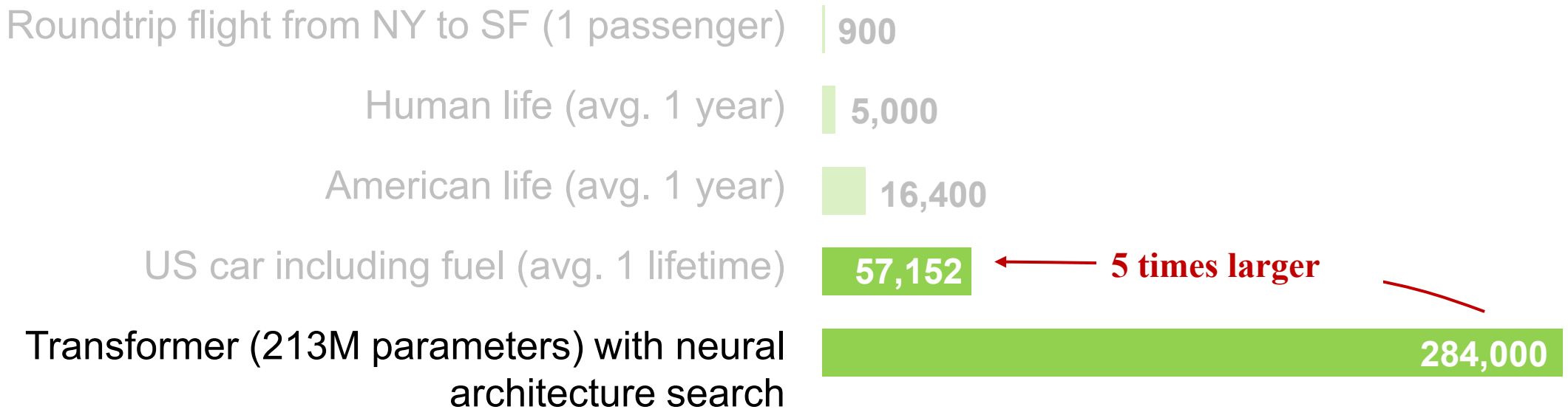


Strubell, *et al*. "Energy and Policy Considerations for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

Background: The Demand for Efficient FL in MEC

find

- **Common carbon footprint benchmarks (in KG of CO2 equivalent)**



Strubell, *et al*. "Energy and Policy Considerations for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

Background: The Demand for Efficient FL in MEC

☐ Deep learning (DL) is rather energy-consuming

- **Exponential Growth of DL Model**: the amount of compute used in the largest AI training runs has been **increasing exponentially with a 3.4-month doubling time** (by comparison, Moore's Law had a 2-year doubling period) [1].
- **Explosive Growth of Data**: over the next several years up to 2025, global data creation is projected to grow to more than 180 zettabytes [2].

☐ Federated learning can be more energy-consuming than centralized training

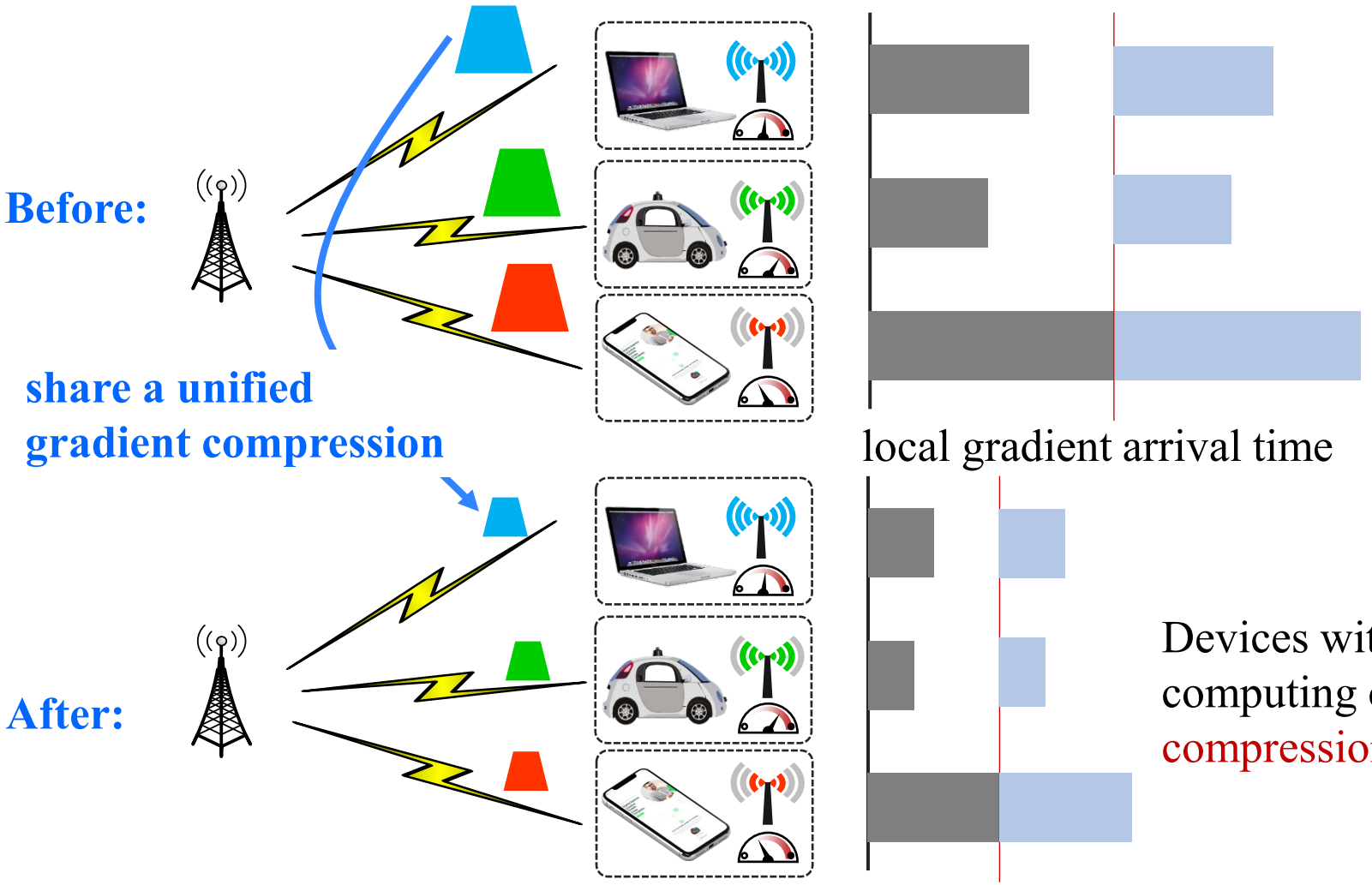
- **Higher communication costs**: many IoT devices connect to the aggregation server via wireless communication, such as 4G/5G, which is highly energy-consuming.
- **More training iteration steps**: due to the non-independent and identically data distribution, federated learning requires many iteration steps to converge.

[1] OpenAI. AI and Compute. https://openai.com/blog/ai-and-compute/
[2] Statista. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/

Existing Energy-Efficient FL Frameworks

find

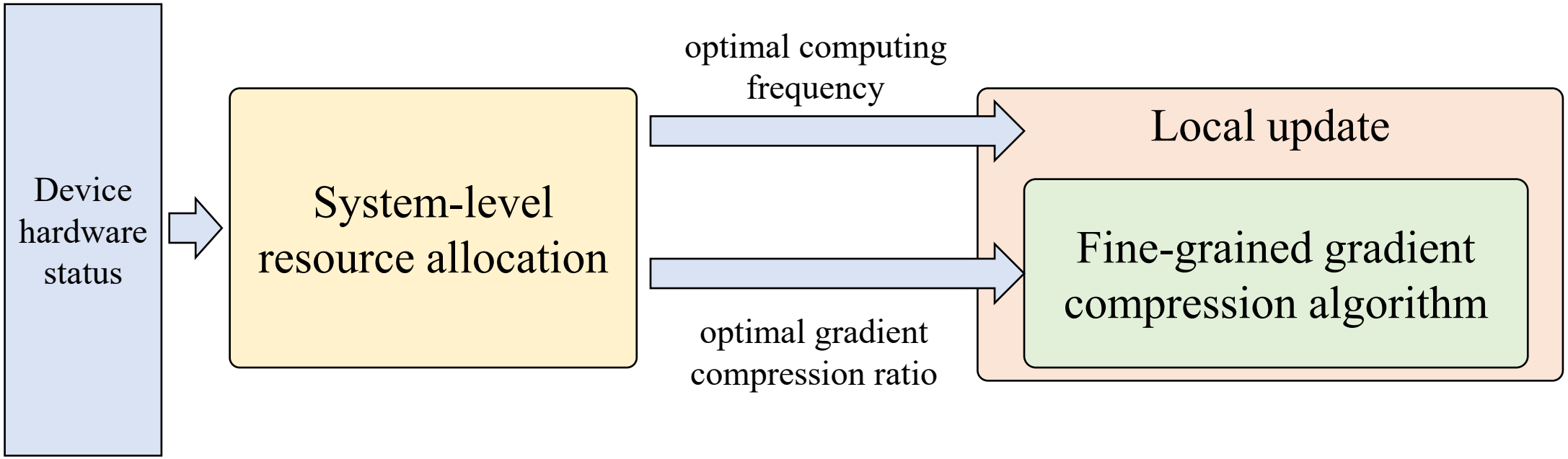
☐ Compression algorithm: gradient sparsification and/or quantization



Devices with various communication states and computing capabilities share a unified gradient compression strategy.

FedGreen: Compression Alg. & Resource Allocation Co-design

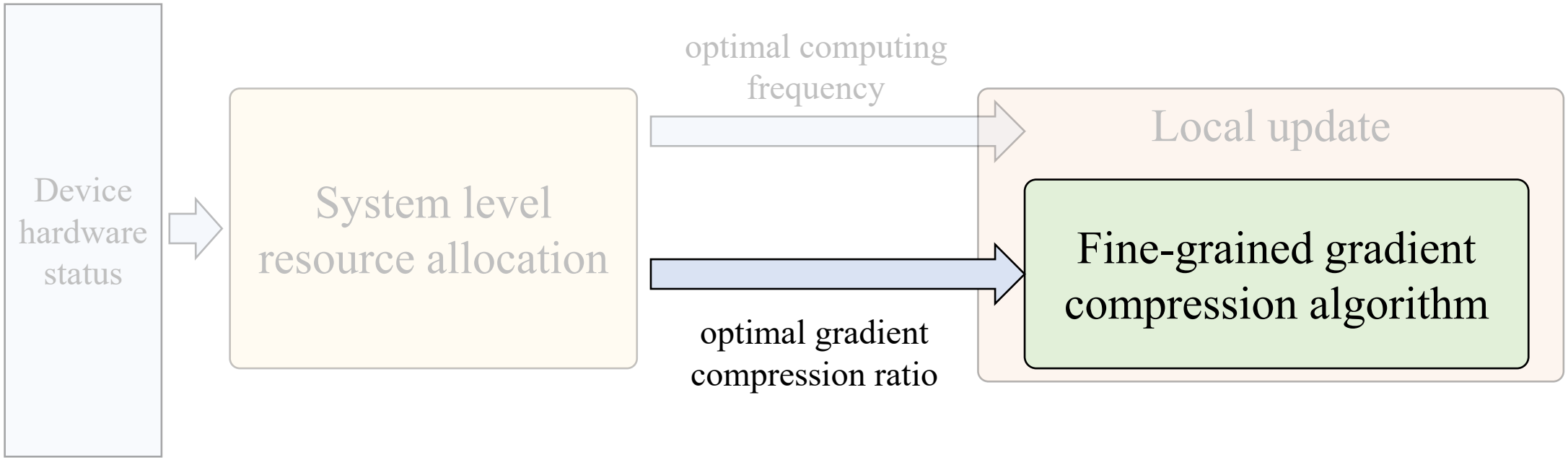
In FL with gradient compression, the computing frequency and compression ratio of each device should **be optimized to match the hardware** configuration and channel status.



FedGreen Part I: Gradient Compression Algorithm

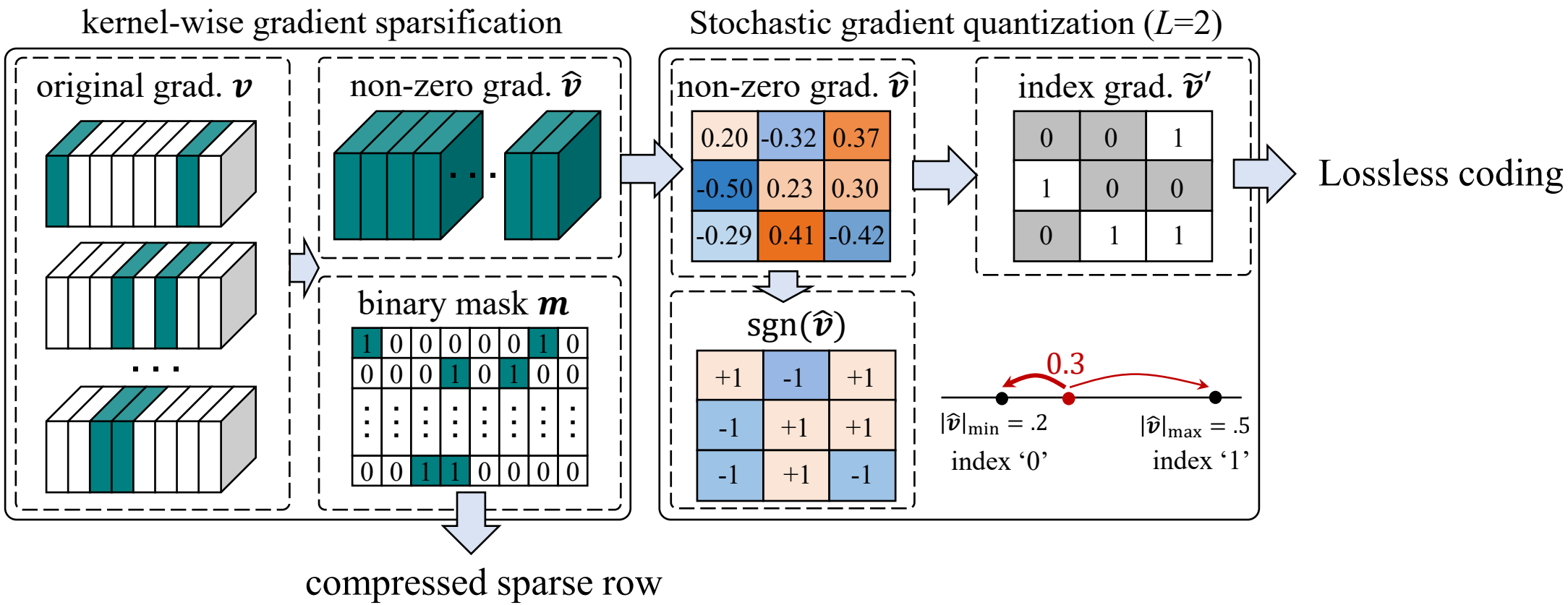
Input: plain local gradient, expected compression ratio

Output: compressed local gradient



Device-Side Fine-Grained Gradient Compression

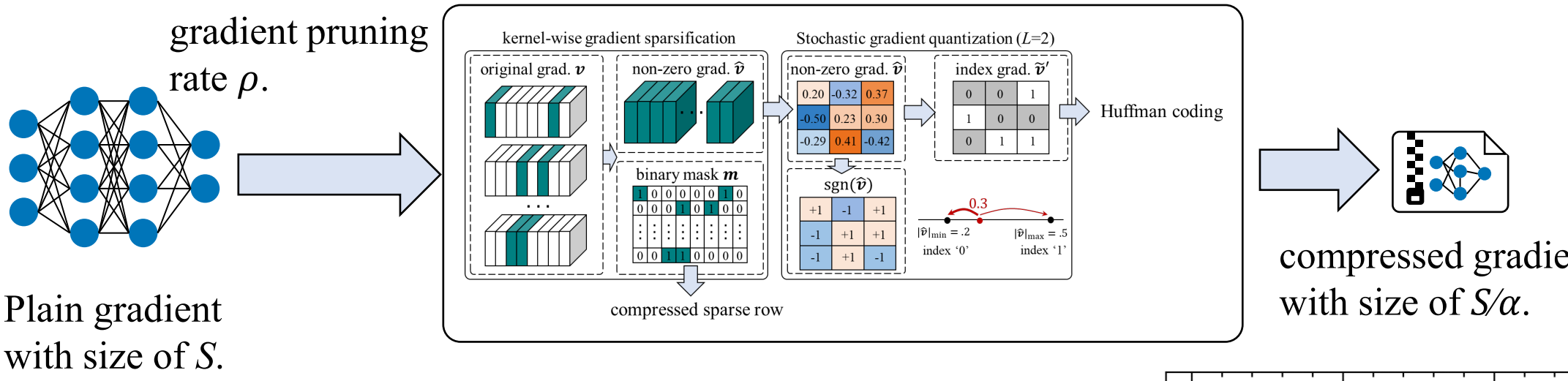
Design a gradient compression algorithm that supports **a continuous range of compression ratios.**



We use a fixed L=8 for convolution layer and L=4 for fully connected layer during the implementation and **the compression ratio is only determined by the gradient sparsity (gradient pruning rate)**.

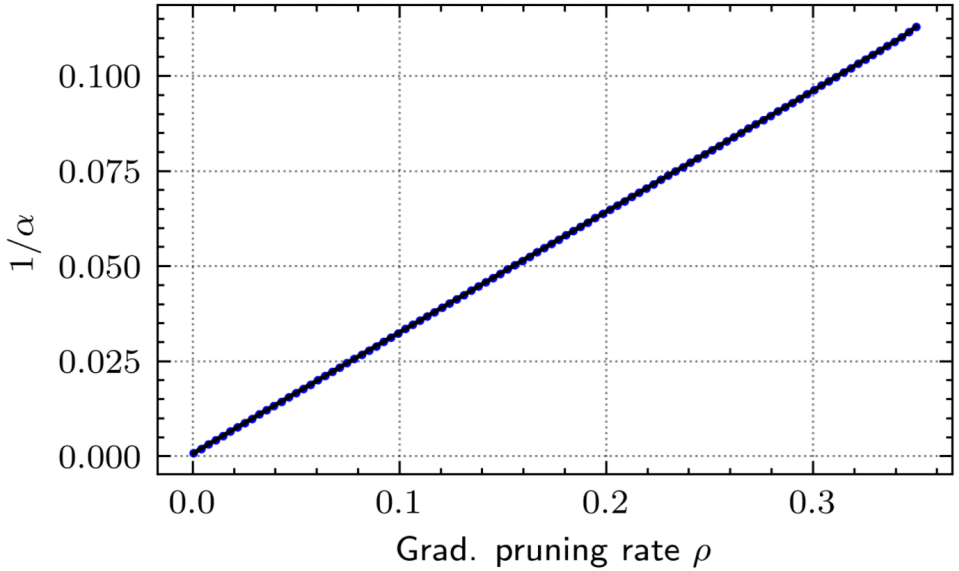
Device-Side Fine-Grained Gradient Compression

Design a gradient compression algorithm that supports **a continuous range of compression ratios.**



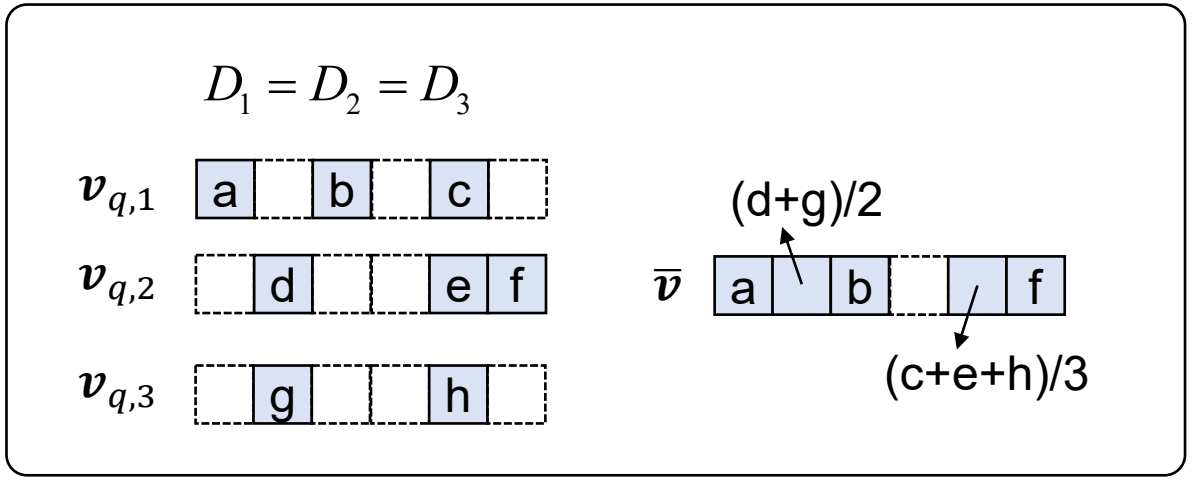
We theoretically show that there is a near-linear relationship between gradient pruning rate ρ and the size of compression gradient S/α.

$$\frac{S}{\alpha} \le \underbrace{C_{out}C_{in}}_{\text{mask size}} + \underbrace{\lceil (1-\rho)C_{out}C_{in} \rceil K^2(1+\log_2 L)}_{\text{size of sgn and non-zero index gradient}} + 64$$



Server-Side Element-wise Aggregation

$$\bar{v}^k = \begin{cases} \dfrac{1}{\sum_i m_i^k D_i} \sum_i v_{q,i}^k m_i^k D_i, & \text{if } \sum_i m_i^k D_i > 0 \\ 0, & \text{otherwise} \end{cases}$$



Compressed Gradient Information

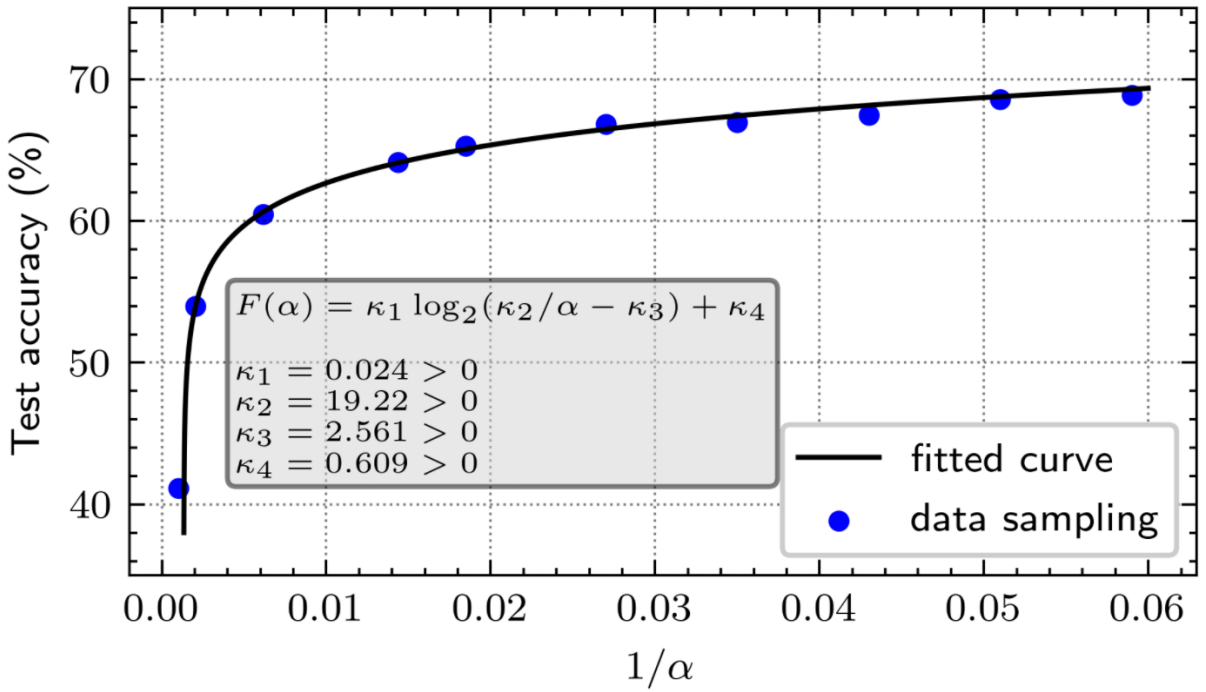
☐ What is the relationship between compression ratio and model accuracy?

We conduct experiments of gradient compression to measure different global model accuracy with respect to different compression ratios of the devices.

We propose to explore the prior knowledge of parameter fitting on a *proxy task with public dataset*, and then transfer it to the *target task with decentralized dataset*. The overall parameter fitting experiments are performed in an offline manner, and the prior knowledge of this *one-time* fitting can be transferred into *many* FL tasks.

Parameter fitting experiment on CINIC dataset:

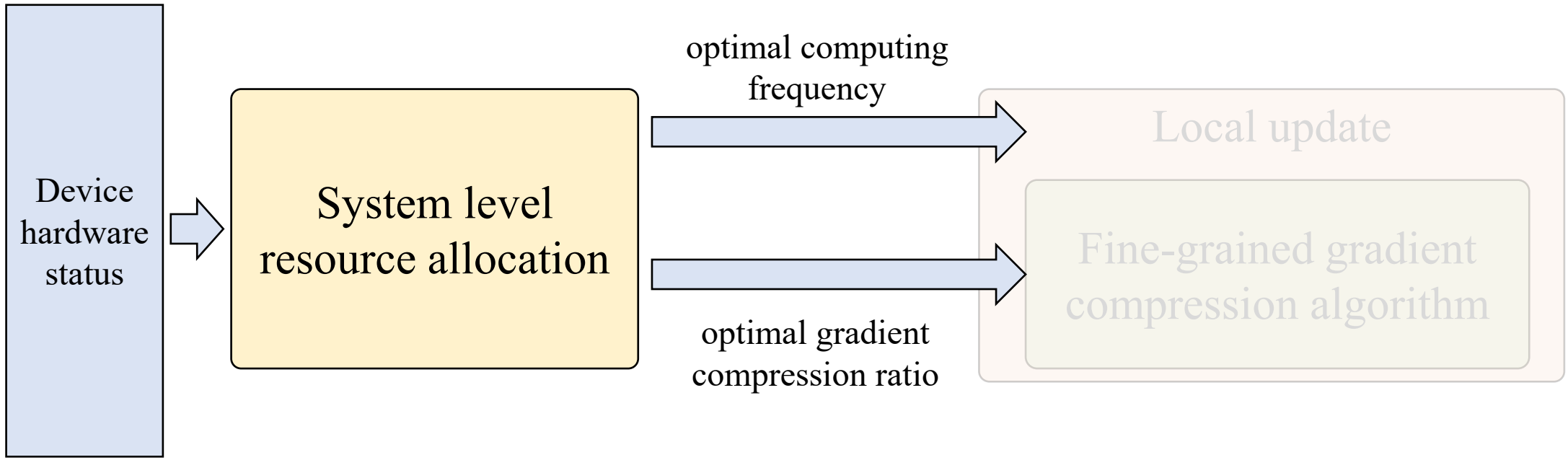
with the decrease of α, more accurate gradient information is collected from the compressed local gradients, which is helpful to improve the global model accuracy.



FedGreen Part II: Resource Allocation

Input: device hardware status

Output: optimal computing frequency, optimal gradient compression ratio



Learning Accuracy-Energy Efficiency Trade-off

find

Communication model:

uplink data rate

$$r_i = b_i \log_2(1 + \frac{p_i |h_i|^2}{N_0 b_i}),$$

time for gradient upload

$$S/(\alpha_i r_i)$$

energy for gradient upload

$$\frac{p_i S}{\alpha_i r_i}$$

Computation model:

time for local training

$$nWD_i/f_i.$$

energy for local training

$$\varepsilon_i f_i^2 n D_i W_i$$

Learning Accuracy-Energy Efficiency Trade-off

The **overall contribution** of all the compressed local gradients from the devices is measured as:

$$\mathcal{F}(\{\alpha_i\}_{1\le i\le I}) = \frac{1}{\mathcal{D}} \sum_{1\le i\le I} D_i F(\alpha_i)$$

Considering the learning performance of FL and total energy consumption of all the devices, there exists a **tradeoff problem** in FL with gradient compression:

learning accuracy energy efficiency

$$\mathcal{G} = F(\{\alpha_i\}_{1\le i\le I}) - \varpi J \sum_{i=1}^{I} \left(\frac{p_i S}{\alpha_i r_i} + \varepsilon_i f_i^2 n D_i W\right)$$

$$\begin{aligned}
&(\mathrm{P1}): && \max \mathcal{G} \\
&\text{subjectto}: && \alpha_i \ge 1, \forall i, \\
& && 0 < f_i \le f_i^{\max}, \forall i, \\
& && \frac{S}{\alpha_i r_i} + \frac{n D_i W}{f_i} \le T^{\max}, \forall i, \\
&\text{variables}: && \alpha_i, f_i, \forall i
\end{aligned}$$

Solution

find

☐ Key idea: substitution method

We utilize an intermediate variable $\beta_i \in [0,1]$ for device i

$$\begin{cases} \frac{S}{\alpha_i r_i} = \beta_i T^{\max}, \\ \frac{nD_iW}{f_i} = (1-\beta_i)T^{\max}. \end{cases}$$

We derive the partial derivatives of $\mathcal{G}$ with respect to β_i

$$\begin{cases} \frac{\partial \mathcal{G}}{\partial \beta_i} = \frac{r_i T^{\max}}{S} \frac{D_i \kappa_1 \kappa_2}{\mathcal{D}(\kappa_2 \lambda_i - \kappa_3)\ln 2} \\ -\varpi J(p_i T^{\max} + 2\varepsilon_i \frac{n^3 {D_i}^3 W^3}{(T^{\max})^2 (1-\beta)^3}), \\ \frac{\partial^2 \mathcal{G}}{\partial \beta_i^2} = -(\frac{r_i T^{\max}}{S})^2 \frac{D_i \kappa_1 \kappa_2^2}{\mathcal{D}(\kappa_2 \lambda_i - \kappa_3)^2 \ln 2} \\ -6\varpi J \frac{\varepsilon_i n^3 {D_i}^3 W^3}{(T^{\max})^2 (1-\beta)^4} < 0, \end{cases}$$

Clearly, $\mathcal{G}$ is concave on β_i and we solve the optimal solution of β_i based on the first-order optimality condition. But it is difficult to directly solve β_i .

Alternatively, we apply the binary search method to seek an approximate solution.

Experiments

☐ Experiment settings

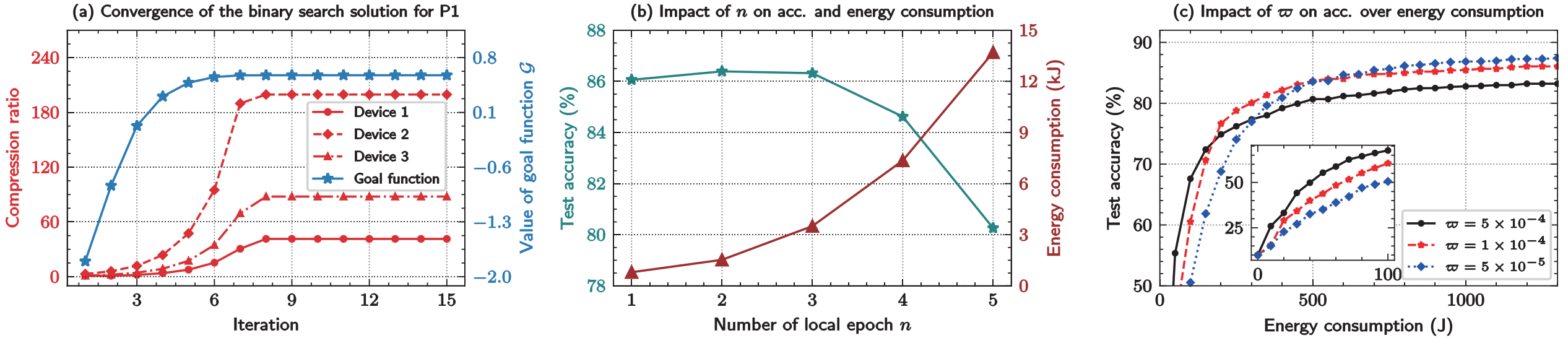
Dataset, model: **CIFAR-10** dataset, with 16 mobile devices, **VGG-9**.

FL hyperparameters: S = 111.7 Mb and W = 0.98 megacycles, local epoch 1, batch size 64, learning rate 0.05, the number of global iterations 300, and decay rate per round 0.996 by default.

Hardware configuration and channel status: energy coefficient $\varepsilon_i \sim U[5 \times 10^{-27}, 1 \times 10^{-26}]$, computing frequency $f_i^{\max} \sim U[1.5, 4]$ GHz, power-spectral-density $N_0 = -114$dBm, available bandwidth $b_i \sim U[0.8, 5]$ MHz. We set the weighting factor $\varpi = 1 \times 10^{-4}$ and $T_{\max} = 100$ seconds by default.

Experiments

☐ The convergence of binary search solution and the impact of different parameters on the overall performance of FedGreen



(a) Convergence of the binary search solution: the binary search method can converge finally and achieve the approximately optimal solution for the trade-off problem.

(b) Impact of local epoch: large local epoch may lead to the divergence of the local gradients, which matches with the existing study.

(c) Impact of weighting factor: large value of ϖ means that the parameter server would like to reduce the total energy consumption of the devices.

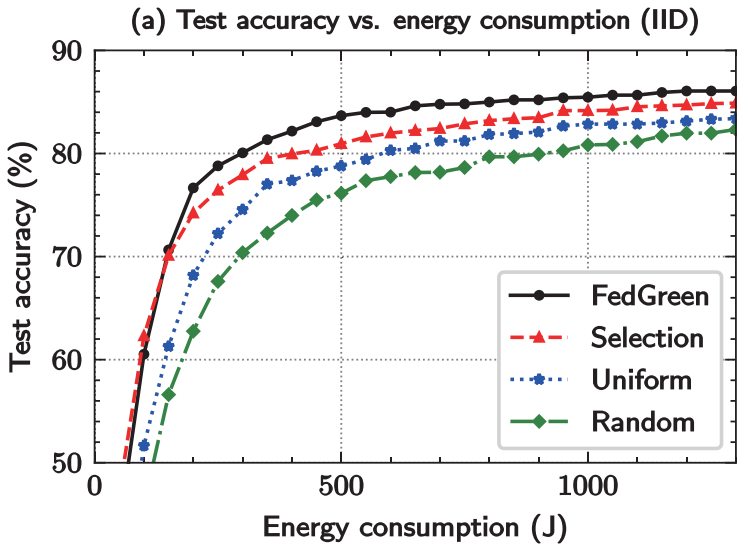
Experiments

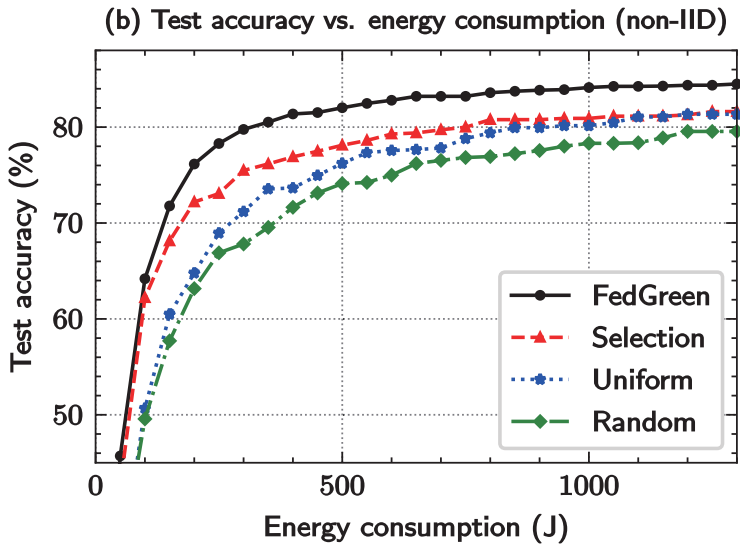
☐ Comparison with baseline

Random: Each device utilizes a random strategy of gradient compression strategy.

Uniform: All the devices utilize an identical ratio for gradient compression. We calculate the average compression ratio $\bar{\alpha}$ of FedGreen.

Selection: exclude the top 25% of the devices with the largest energy consumption in the uniform policy.





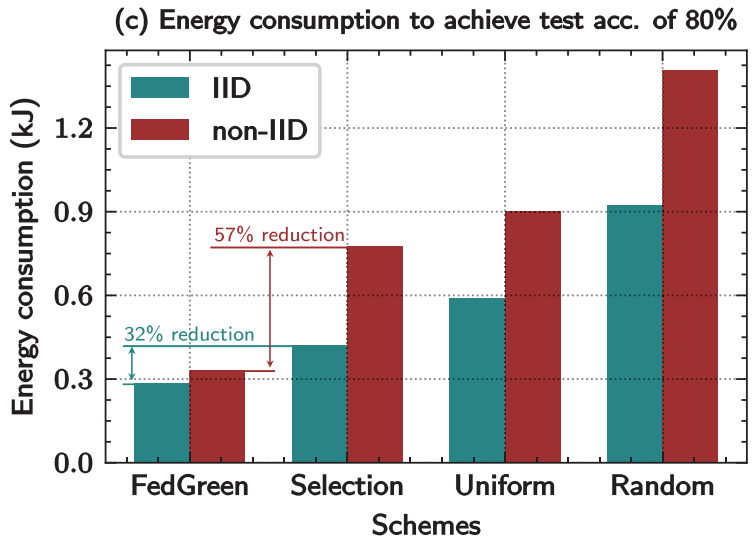


Fig. (a) and Fig. (b): with the same energy consumption requirement, our scheme outperforms the existing schemes to improve the global model accuracy.

Fig. (c): to achieve the same test accuracy of 80%, compared with the Selection scheme, FedGreen reduces 32% and 57% of the energy consumption under the IID and non-IID setting, respectively.

THANKS

Project page http://fedgreen.find-lab.com/