Random Orthogonalization Design

Xizixiang Wei

[C] Xizixiang Wei, Cong Shen, Jing Yang and H. Vincent Poor, Random Orthogonalization for Federated Learning in Massive MIMO Systems, in Proc. IEEE International Conference on Communications (ICC), May 2022.

[J] Xizixiang Wei, Cong Shen, Jing Yang and H. Vincent Poor, Random Orthogonalization for Federated Learning in Massive MIMO Systems, IEEE Transactions on Wireless Communications, 2023.

Background: Federated Learning

1

Machine learning enables powerful applications

2

Massive real-word data are generated at edge devices

3

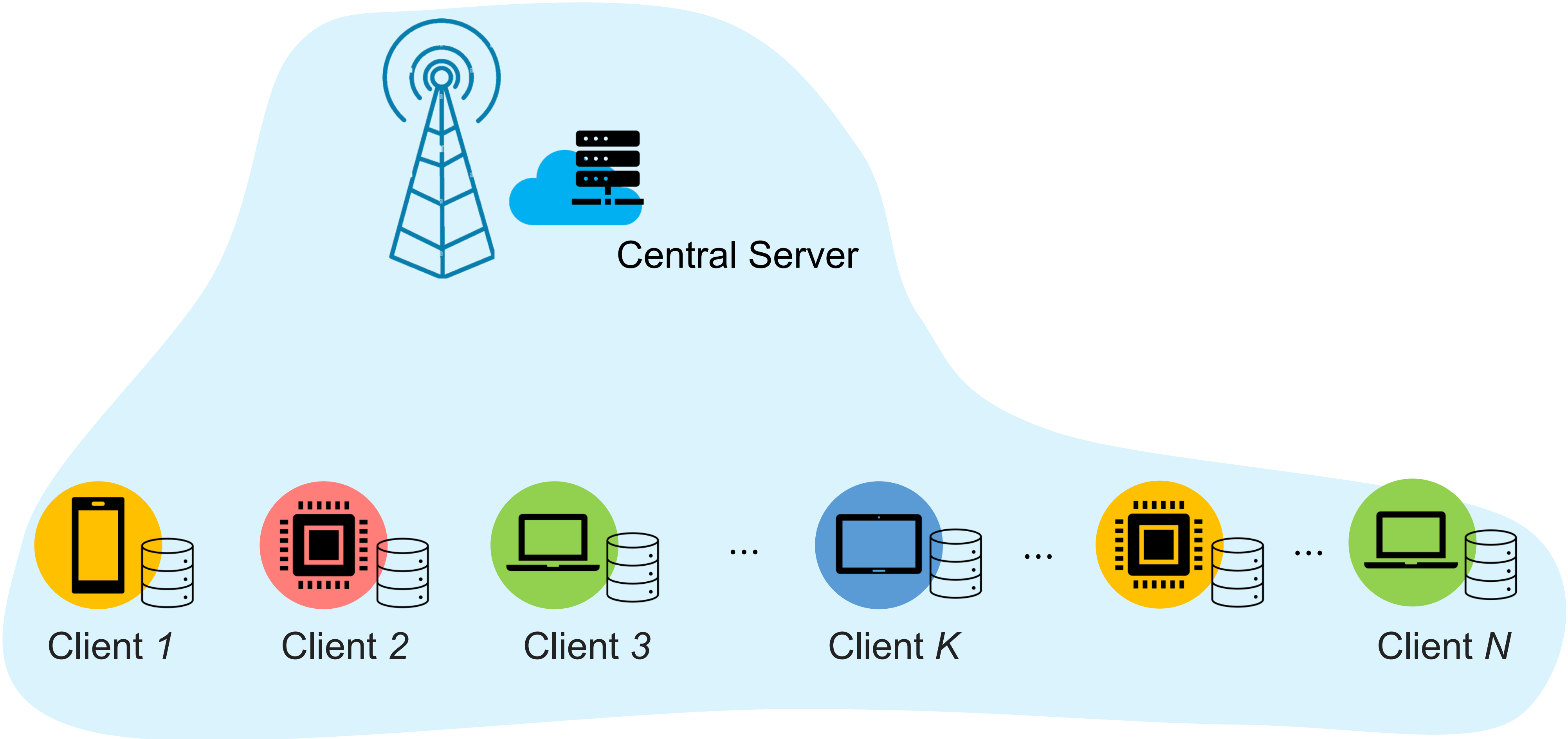
We want to keep sensitive data at edge devices

Federated Learning (FL)

- A distributed machine learning paradigm
- Obtain a global model while keeping data locally

FedAvg

Goal: central server obtain a **global model** trained by **local data** at total N clients.

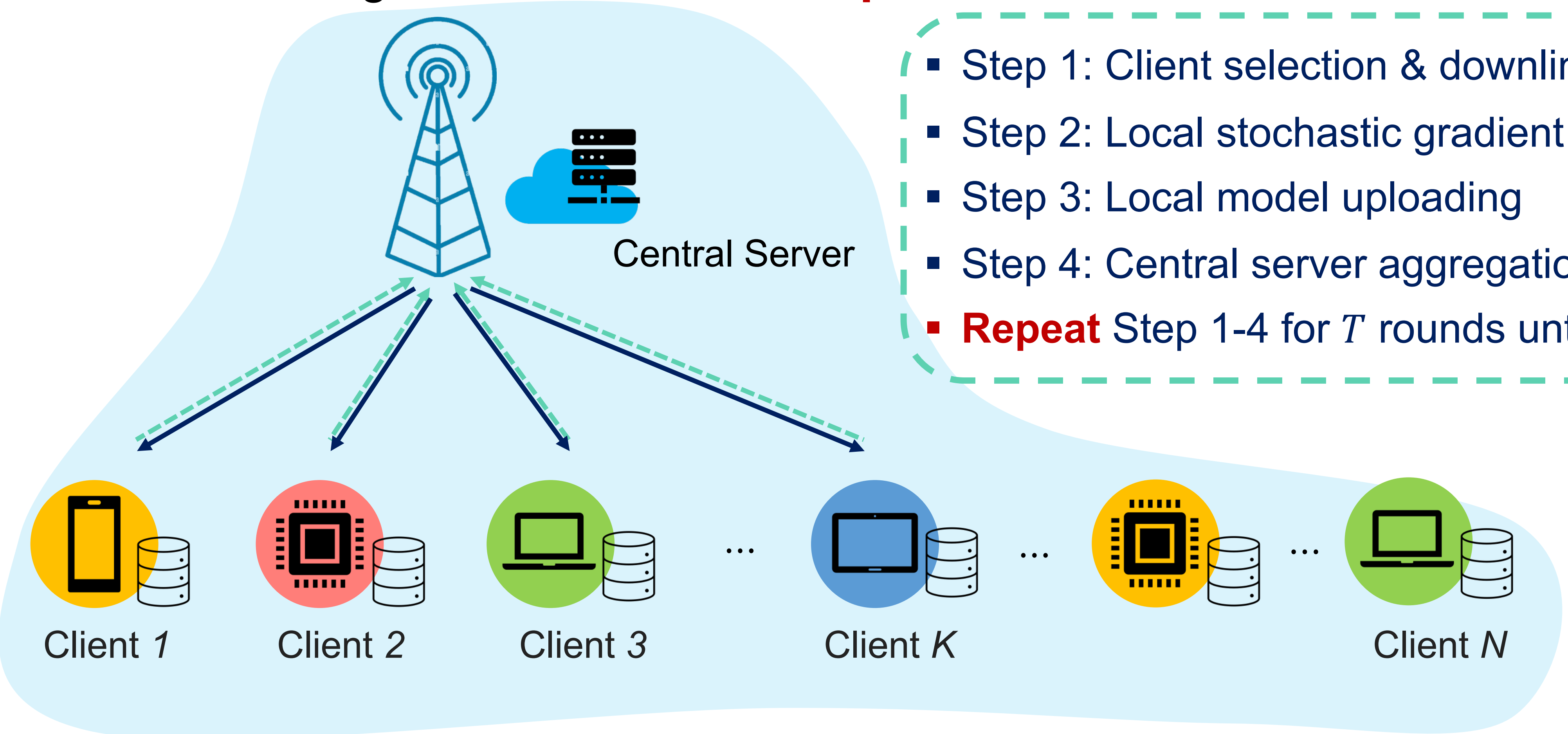


Ref: McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

FedAvg

FedAvg: a composition of **multiple learning rounds**

Each learning round contains **4 steps**.

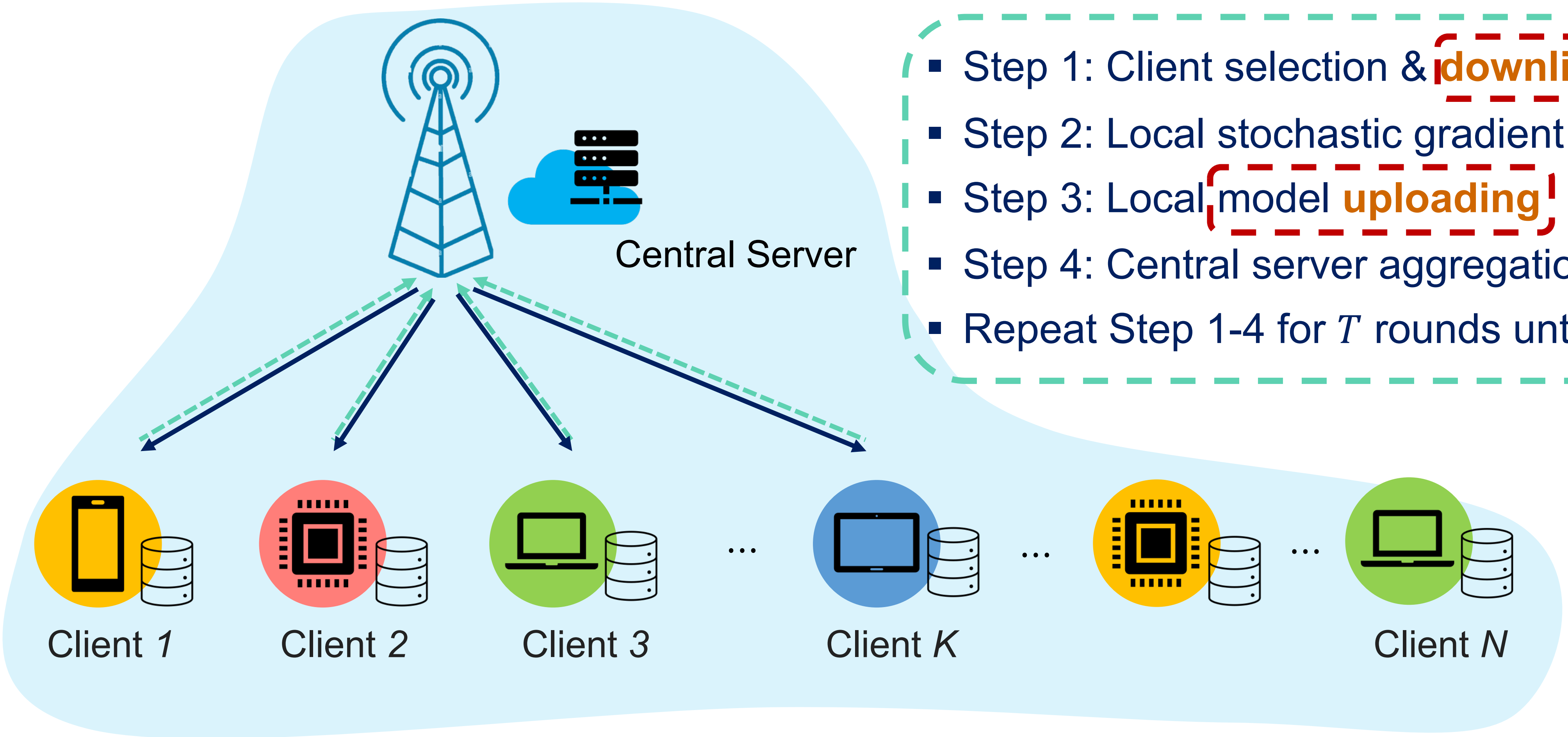


- Step 1: Client selection & downlink broadcast
- Step 2: Local stochastic gradient descent (SGD)
- Step 3: Local model uploading
- Step 4: Central server aggregation
- **Repeat** Step 1-4 for T rounds until **convergence**

Ref: McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

FedAvg

Communication is the **bottleneck** of federated learning.



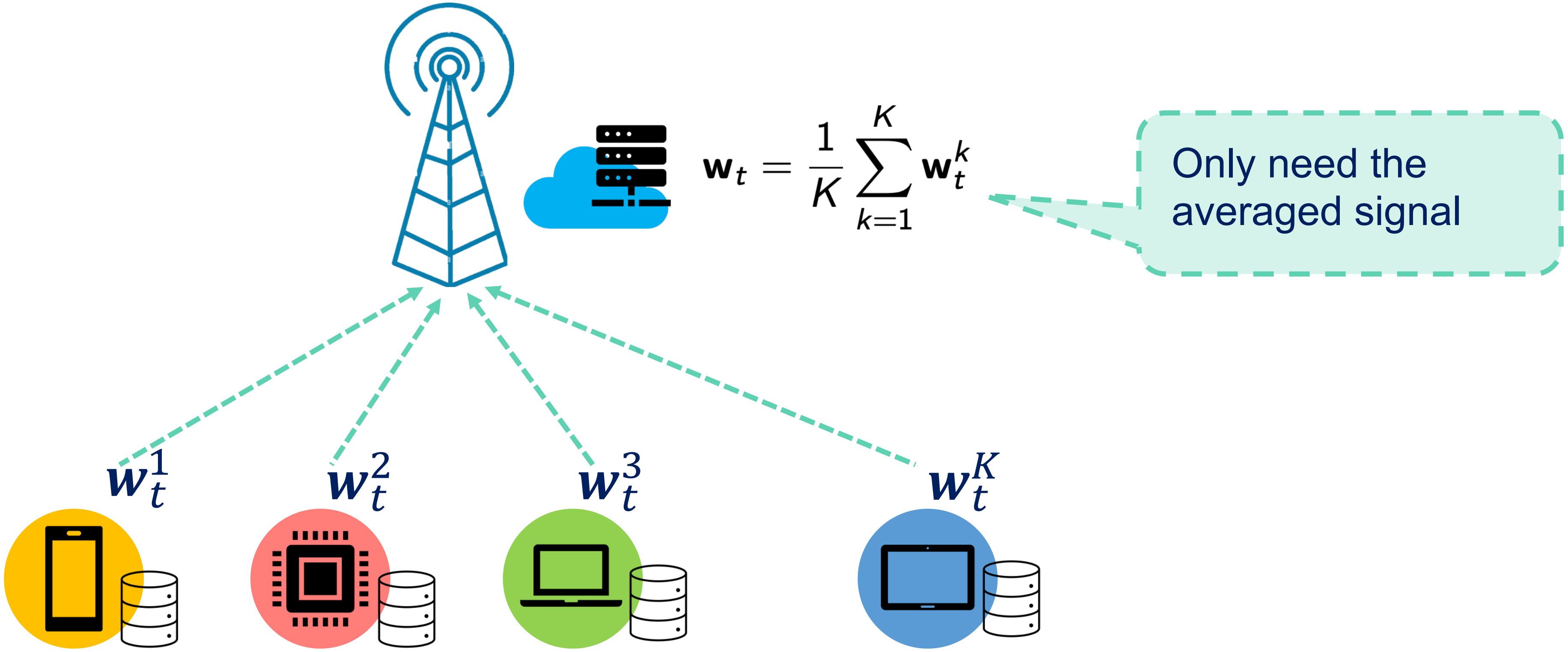
- Step 1: Client selection & **downlink** broadcast
- Step 2: Local stochastic gradient descent (SGD)
- Step 3: Local model **uploading**
- Step 4: Central server aggregation
- Repeat Step 1-4 for T rounds until convergence

Ref: McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

Motivation

Main difference from traditional communications

- Central server does **not** need to decode **individual model** in uplink comm.



Motivation

Main difference from traditional communications

- Central server does **not** need to decode **individual model** in uplink comm.

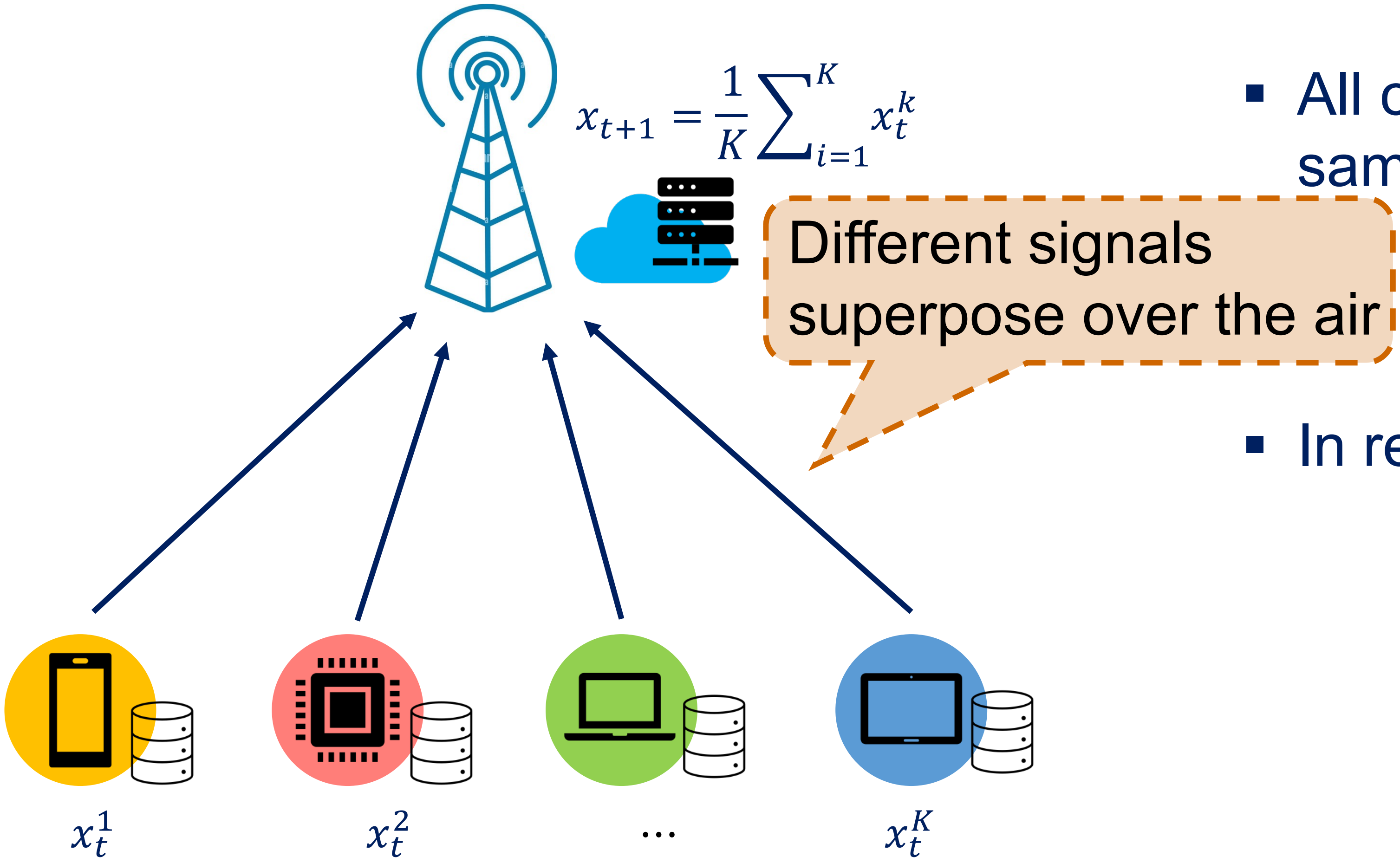
Characteristics of federated learning

Properties of wireless channel

Co-design more efficient and reliable wireless communication schemes for federated learning

Uplink of FL: scaling challenges

Over-the-Air Computation **(AirComp)** is a promising solution.



- All clients can be scheduled at the same time-frequency resource

$$x_{t+1} = y = \sum_{i=1}^{K} x_t^k$$

- In real wireless channel

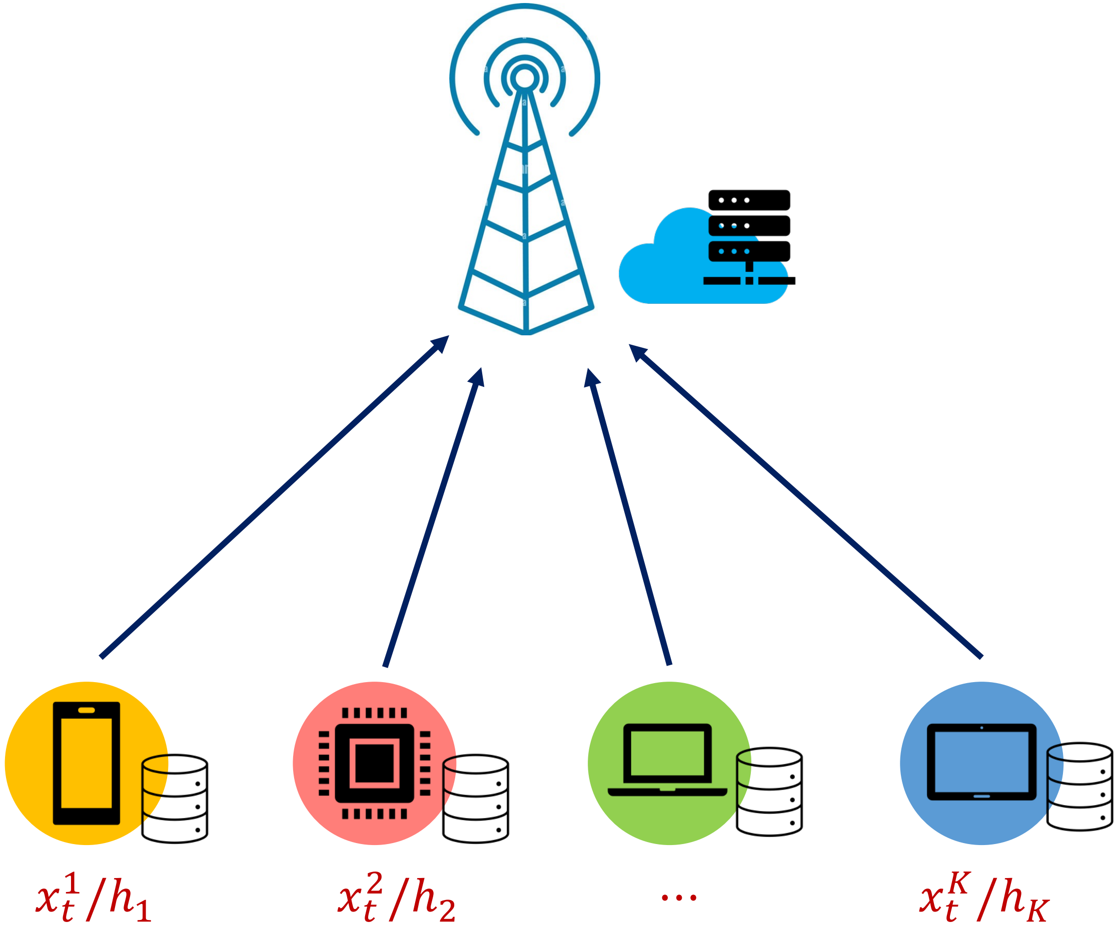
$$\hat{x}_{t+1} = y = \sum_{i=1}^{K} h_k x_t^k + n$$

Biased estimation

Ref: Guangxu Zhu, Yong Wang, and Kaibin Huang. "*Broadband analog aggregation for low-latency federated edge learning.*" IEEE Transactions on Wireless Communications, 2019.

Uplink of FL: AirComp

Heuristic channel inversion



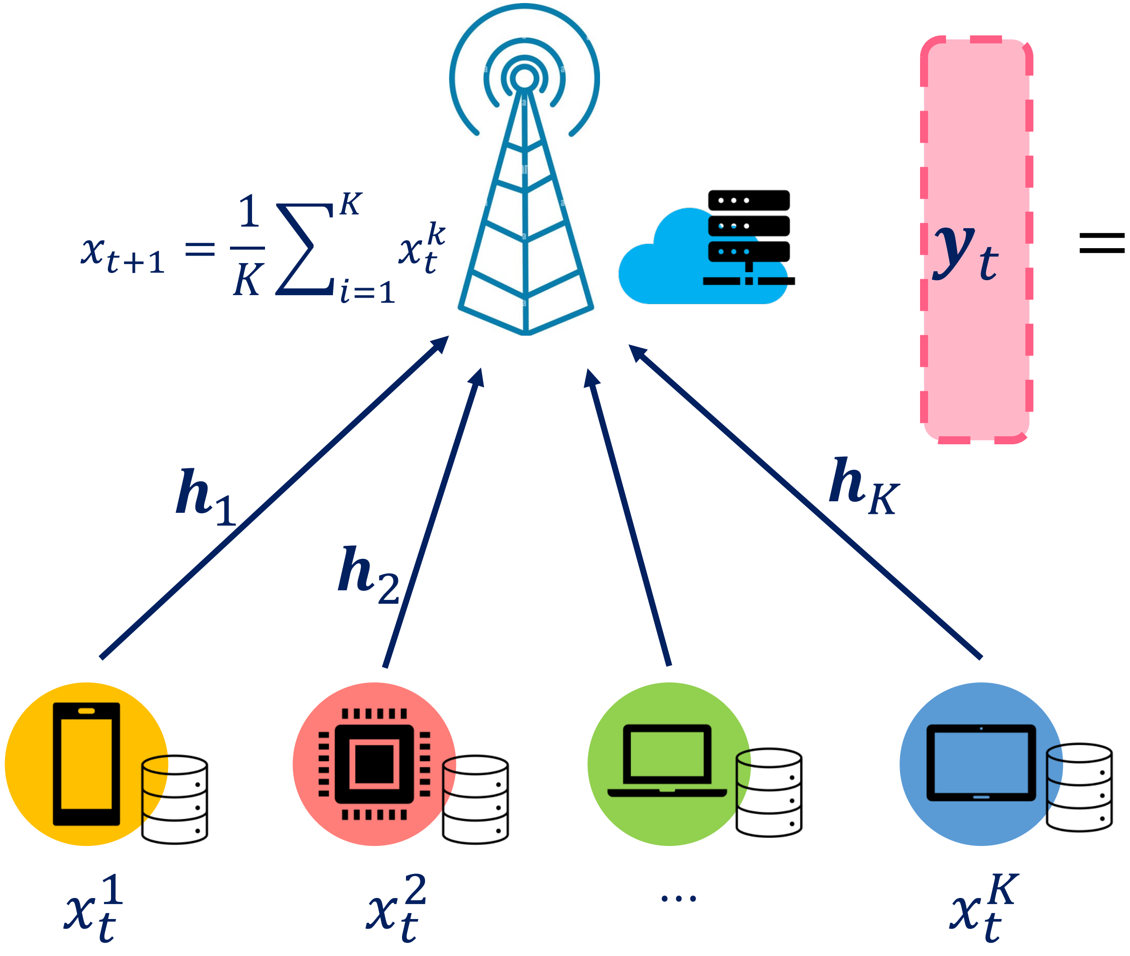
$$\hat{x}_{t+1} = y = \sum_{i=1}^{K} h_k \frac{x_t^k}{h_k} + n = \sum_{i=1}^{K} x_t^k + n$$

- Require channel state information at transmitters (CSIT)
- Increasing dynamics of signal
- Performance will “blow up” when deep fading

Ref: Guangxu Zhu, Yong Wang, and Kaibin Huang. "*Broadband analog aggregation for low-latency federated edge learning.*" IEEE Transactions on Wireless Communications, 2019.

Uplink of FL: AirComp + MIMO

Solution: Using **high-dimension** $\boldsymbol{h}_k \in \mathbb{C}^{M\times 1}$ provided by **massive MIMO**



$$\boldsymbol{y}_t = \boldsymbol{h}_1 x_t^1 + \boldsymbol{h}_2 x_t^2 + \cdots + \boldsymbol{h}_K x_t^K + \boldsymbol{n}_t$$

Goal: Find **a linear projector**

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{v}^H \mathbf{y}_t$$

Channel Hardening and Favorable Propagation

IID Rayleigh fading channel model $\boldsymbol{h}_k \sim CN(0, \frac{1}{M}\boldsymbol{I})$

Channel hardening

$$\mathbf{h}_k^H \mathbf{h}_k \to 1, \quad \text{as} \quad M \to \infty.$$

Favorable propagation

$$\mathbf{h}_k^H \mathbf{h}_j \to 0, \quad \text{as} \quad M \to \infty, \quad \forall k \neq j.$$

Massive MIMO $\xrightarrow{M \longrightarrow \infty}$ **Random Orthogonalization**

Linear projector: sum channel

$$\mathbf{v} = \mathbf{h}_s = \sum_{k=1}^{K} \mathbf{h}_k$$

Random Orthogonalization

Linear projection $\boldsymbol{h}_S^H \boldsymbol{y}$: **an unbiased estimation**

$$\tilde{x}_i = \mathbf{h}_s^H \mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i$$

$$\overset{(a)}{=} \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\text{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i}}_{\text{Interference}} + \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i}_{\text{Noise}}$$

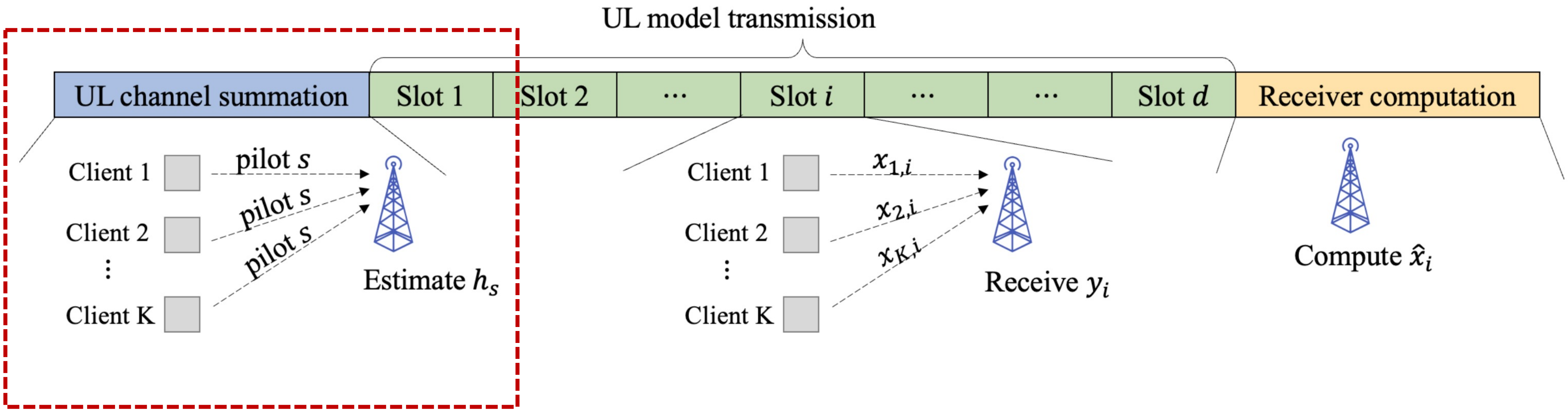
$$\overset{(b)}{\approx} \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d.$$

$\boldsymbol{h}_S^H \boldsymbol{y}$ is an unbiased estimator of sum signal

$$\mathbf{h}_k^H \mathbf{h}_k \to 1, \quad \text{as} \quad M \to \infty.$$

$$\mathbf{h}_k^H \mathbf{h}_j \to 0, \quad \text{as} \quad M \to \infty,$$

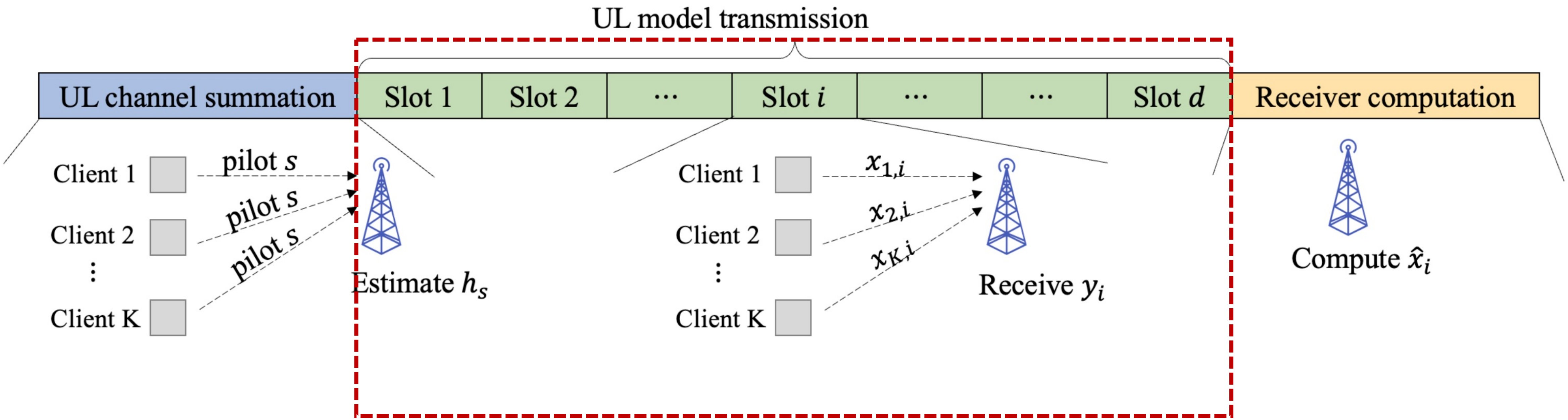
RO: UL Design



$$\mathbf{y}_s = \sum_{k=1}^{K} \mathbf{h}_k s + \mathbf{n}_s \xrightarrow{\text{Estimate}} \mathbf{h}_s = \sum_{k=1}^{K} \mathbf{h}_k$$

- **Partial** CSI at the receiver **(CSIR)**
- Low communication overhead

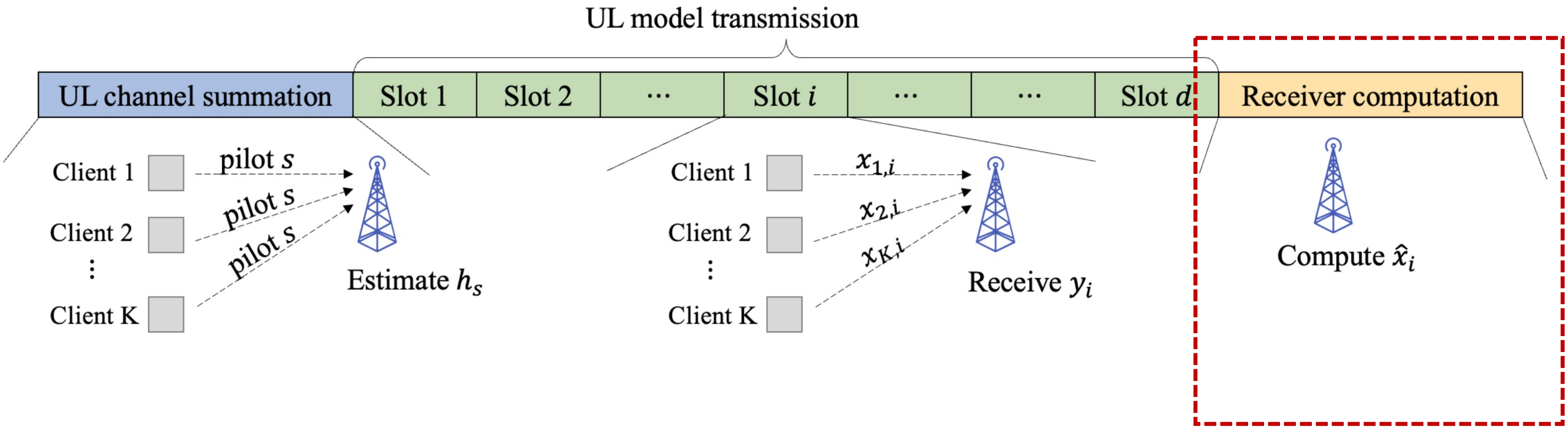
RO: UL Design



$$\boldsymbol{y}_t = \boldsymbol{h}_1 x_t^1 + \boldsymbol{h}_2 x_t^2 + \cdots + \boldsymbol{h}_K x_t^K + \boldsymbol{n}_t$$

- **No CSIT required**

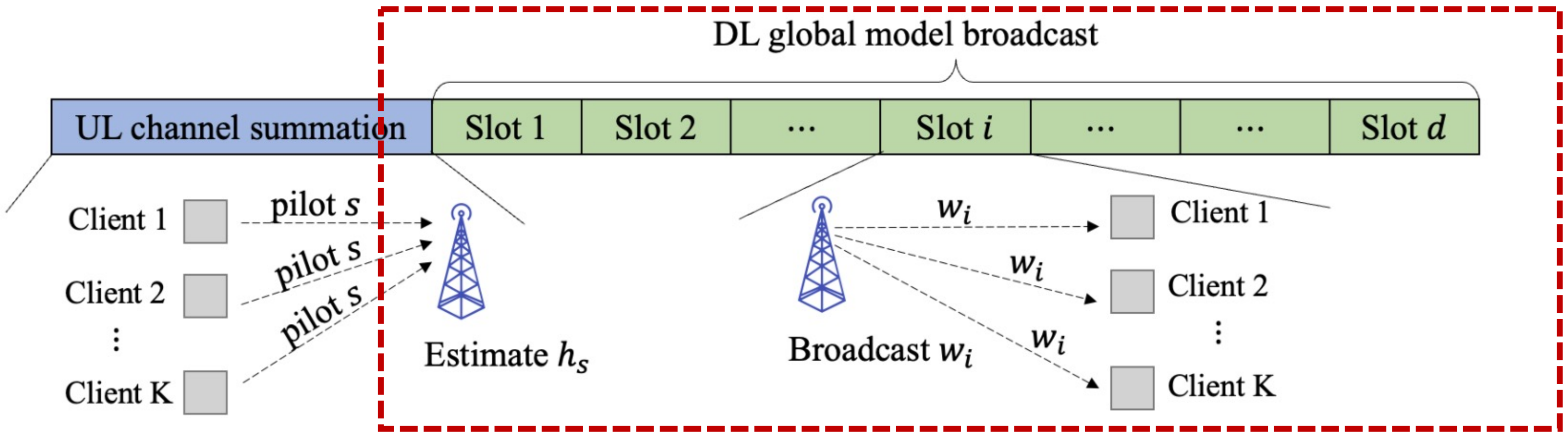
RO: UL Design



$$\tilde{\mathbf{x}}_{t+1} = \mathbf{v}^H \mathbf{y}_t$$

- **Low computational complexity**

RO: DL Design



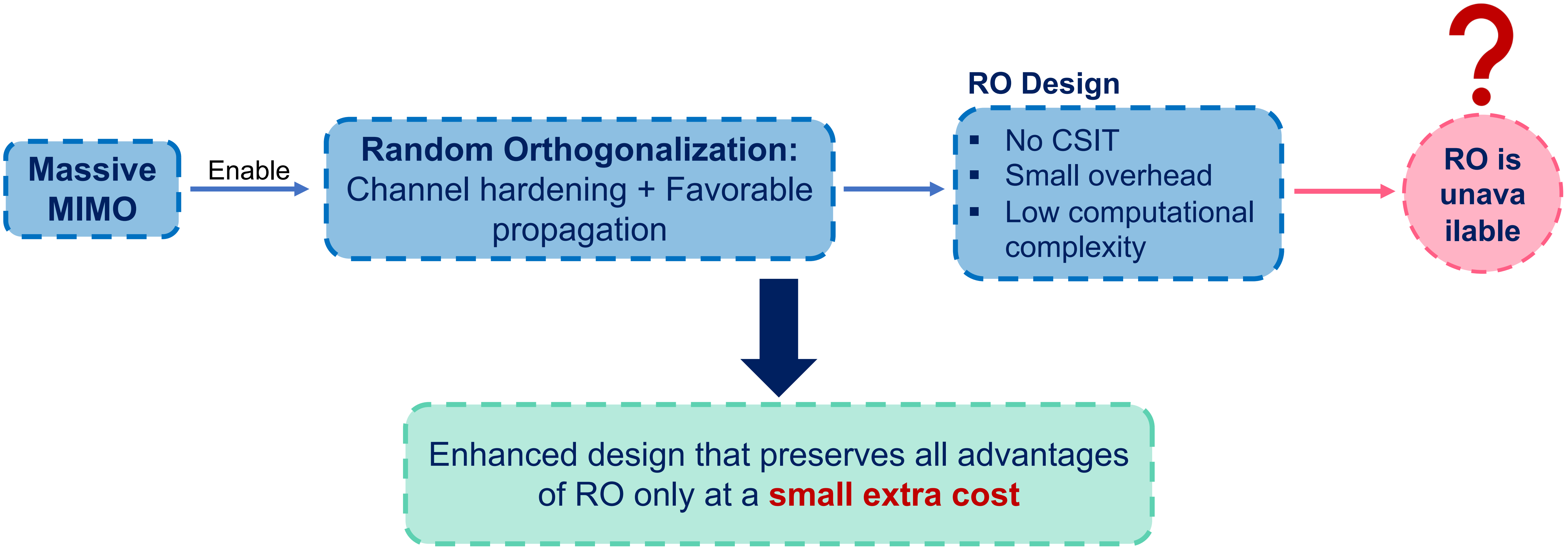
Using $\boldsymbol{h}_s$ as the precoder: an efficient broadcast scheme

$$y_k = \mathbf{h}_k^H \mathbf{h}_s w_i + z_i^k \overset{(a)}{=} \underbrace{\mathbf{h}_k^H \mathbf{h}_k w_i}_{\text{Signal}} + \underbrace{\sum_{j \in [K], j \neq k} \mathbf{h}_k^H \mathbf{h}_j w_i}_{\text{Interference}} + \underbrace{z_i^k}_{\text{Noise}} \overset{(b)}{\approx} w_i \quad \forall i = 1, \cdots, d.$$

$$\mathbf{h}_k^H \mathbf{h}_k \to 1, \quad \text{as} \quad M \to \infty.$$

$$\mathbf{h}_k^H \mathbf{h}_j \to 0, \quad \text{as} \quad M \to \infty,$$

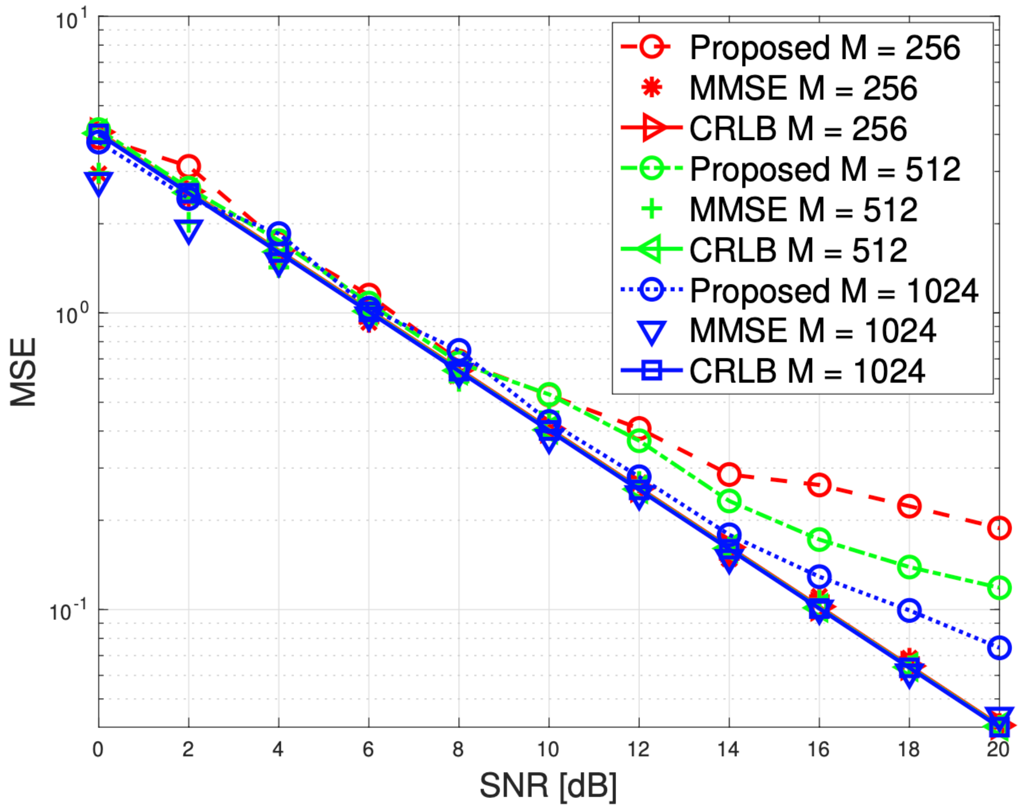
Summary and Enhanced Design



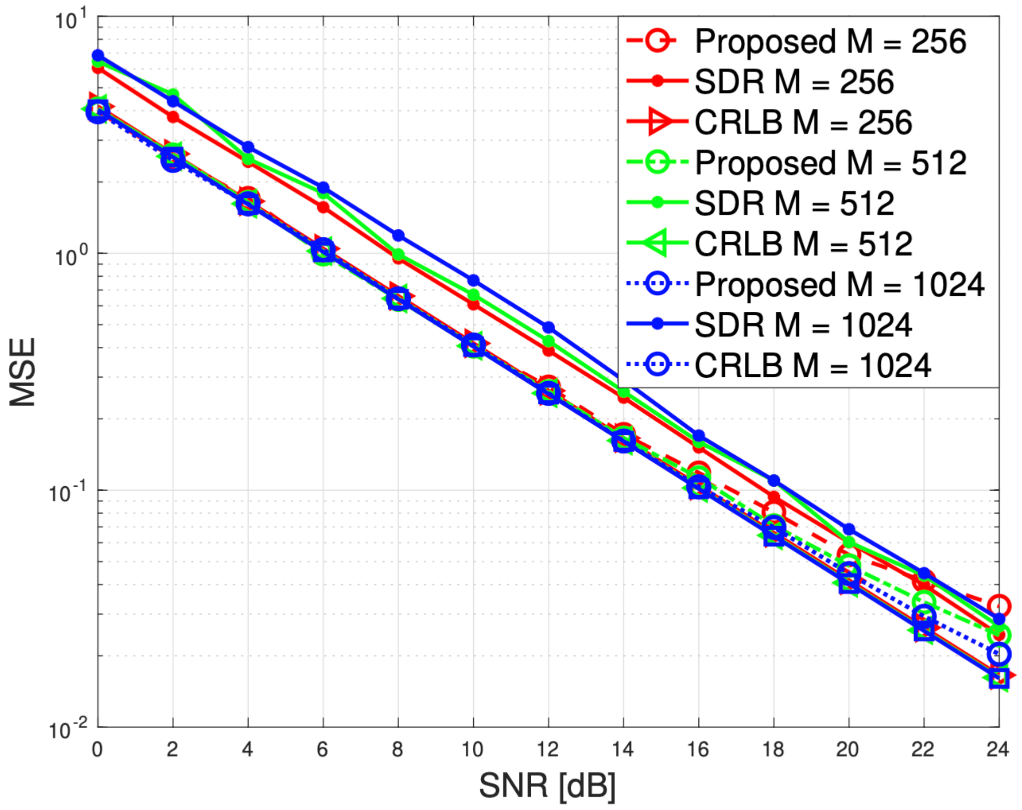
Performance

- Communication performance

Uplink

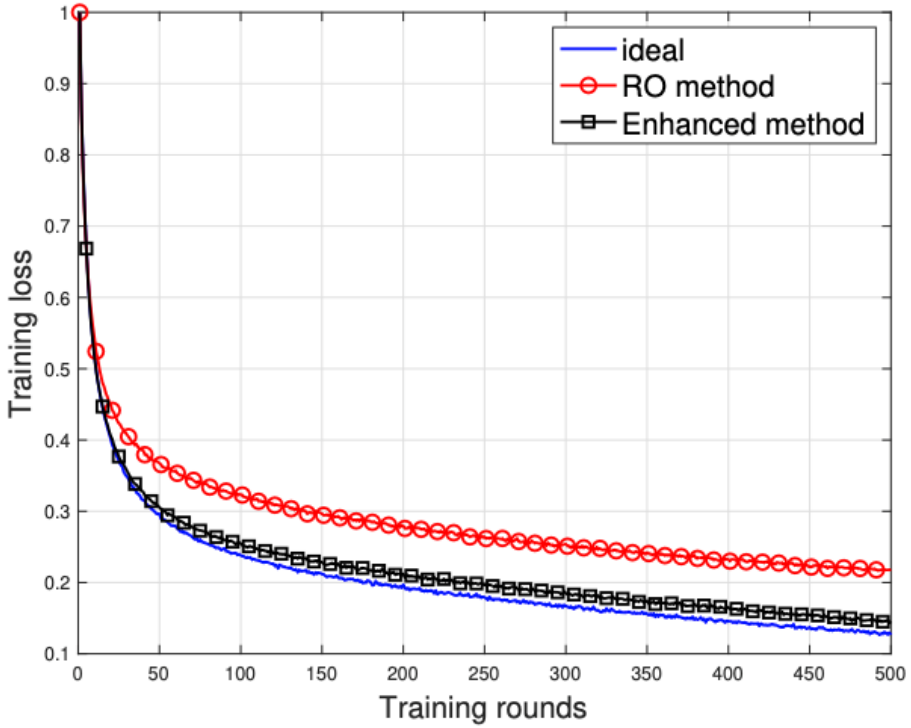


Downlink

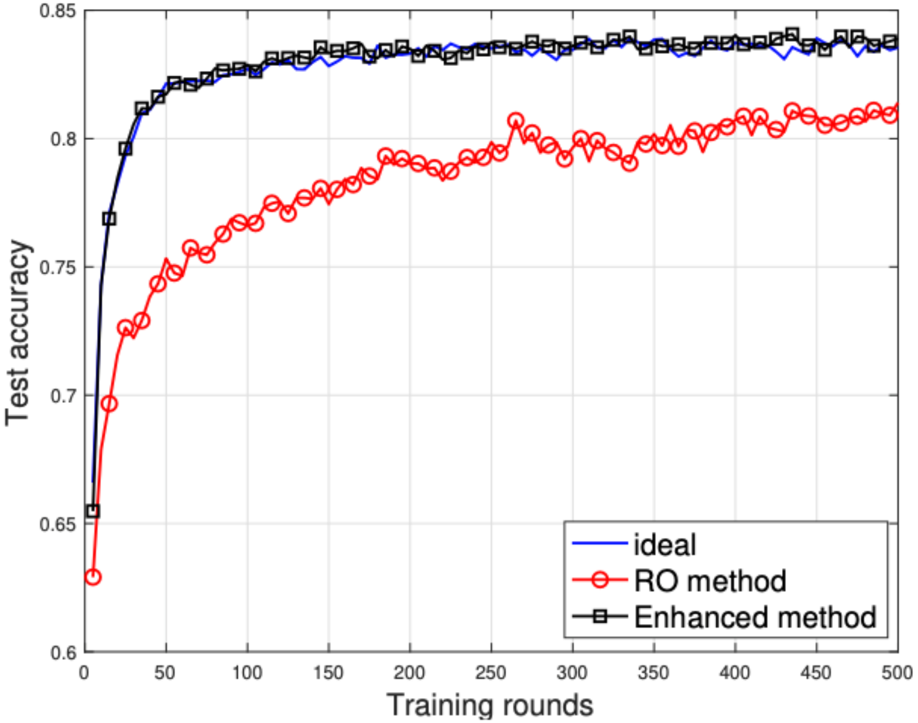


Performance

- Learning performance



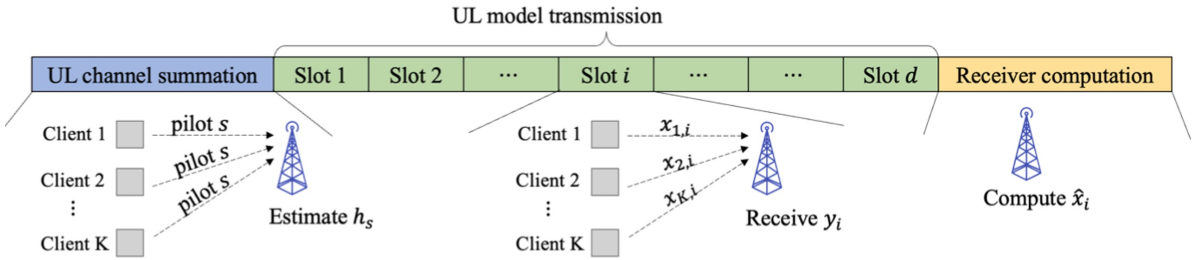
(e) CIFAR-10 uplink+downlink



(f) CIFAR-10 uplink+downlink

Backup

Random Orthogonalization: Uplink Design



- ## Step 1: Uplink channel summation

All clients transmit a common pilot signal s synchronously. The received signal at the BS is

$$\mathbf{y}_s = \sum_{k\in[K]} \mathbf{h}_k s + \mathbf{n}_s,$$

so that the BS can estimate the summation channel $\mathbf{h}_s \triangleq \sum_{k\in[K]} \mathbf{h}_k$

- ## Step 2: Uplink model transmission

All clients transmit model differential parameters to the BS in d shared time slots.

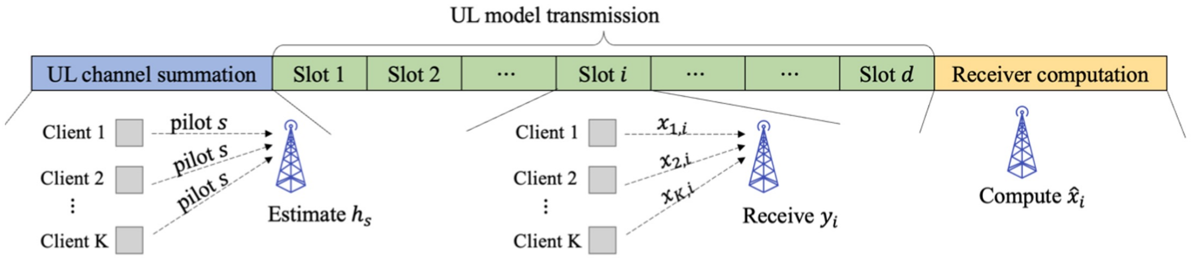
$$\mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \quad \forall i = 1, \cdots, d.$$

- ## Step 3: Receiver computation

The BS estimates aggregated parameter via a simple **linear projection** operation:

$$\tilde{x}_i = \mathbf{h}_s^H \mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i$$

Random Orthogonalization: Uplink Design



- Linear projection: **an unbiased estimation**

$$\tilde{x}_i = \mathbf{h}_s^H \mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i$$

$$\stackrel{(a)}{=} \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\text{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K],j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i}}_{\text{Interference}} + \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i}_{\text{Noise}}$$

$$\stackrel{(b)}{\approx} \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d.$$

Advantages:

Only require patrial CSIT

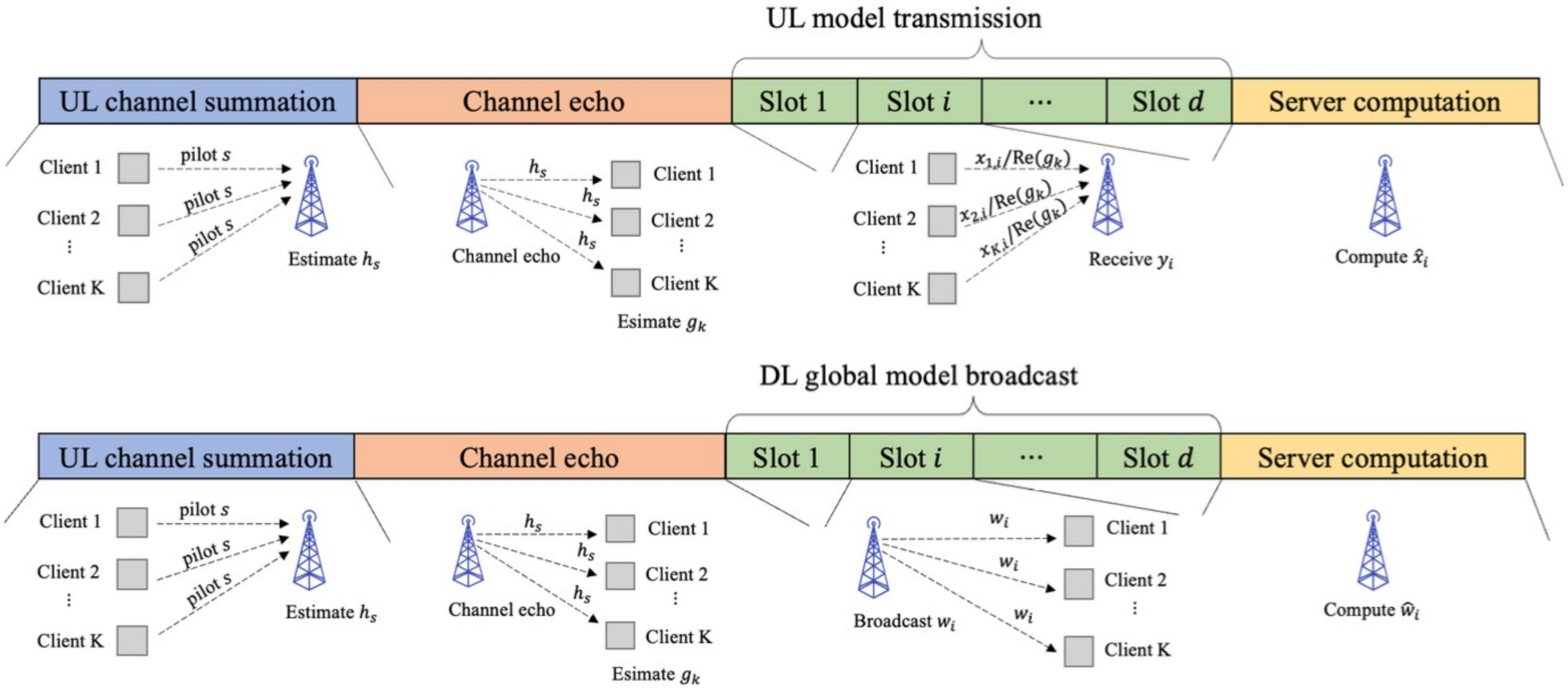
Extremely low complexity

No individual parameter decoded

$$\mathbf{h}_k^H \mathbf{h}_k \to 1, \quad \text{as} \quad M \to \infty.$$

$$\mathbf{h}_k^H \mathbf{h}_j \to 0, \quad \text{as} \quad M \to \infty,$$

Enhanced Design



Channel echo: $g_k = \mathbf{h}_k^H \mathbf{h}_s$

Convergence Analysis

Assumption 1. *L-smooth:* $\forall$ $\mathbf{v}$ *and* $\mathbf{w}$, $\|f_k(\mathbf{v}) - f_k(\mathbf{w})\| \le L\|\mathbf{v} - \mathbf{w}\|$;

Assumption 2. *μ-strongly convex:* $\forall$ $\mathbf{v}$ *and* $\mathbf{w}$, $\langle f_k(\mathbf{v}) - f_k(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle \ge \mu\|\mathbf{v} - \mathbf{w}\|^2$;

Assumption 3. ***Bounded variance for unbiased mini-batch SGD:*** $\forall k \in [N]$,

$$\mathbb{E}[\nabla\tilde{f}_k(\mathbf{w})] = \nabla f_k(\mathbf{w}) \quad \textit{and} \quad \mathbb{E}\left\|\nabla f_k(\mathbf{w}) - \nabla\tilde{f}_k(\mathbf{w})\right\|^2 \le H_k^2;$$

Assumption 4. ***Uniformly bounded gradient:*** $\forall k \in [N]$, $\mathbb{E}\left\|\nabla\tilde{f}_k(\mathbf{w})\right\|^2 \le H^2$ *for all mini-batch data.*

Preserve $O\left(\frac{1}{T}\right)$ convergence rate of SGD

Theorem 1 (***Convergence for random orthogonalization***). *With Assumptions 1-4, for some* $\gamma \ge 0$, *if we select the learning rate as* $\eta_t = \frac{2}{\mu(t+\gamma)}$, *we have*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \le \frac{L}{2(t+\gamma)}\left[\frac{4B}{\mu^2} + (1+\gamma)\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right], \tag{14}$$

for any $t \ge 1$, *where*

$$B \triangleq \left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}}\right]\frac{H^2}{K}. \tag{15}$$