

Latency Minimization in AirComp-Assisted Split Federated Learning with Layer-Wise Aggregation

Haowen Sun^{1,5}, Ming Chen^{1,2}, Yuntao Hu³, Yihan Cang⁴, Quoc-Viet Pham⁵

¹Southeast University, China · ²Purple Mountain Laboratories · ³China Mobile (Hangzhou)

⁴The University of Hong Kong · ⁵Trinity College Dublin, Ireland

Presenter: **Haowen Sun**

28 May, 2026

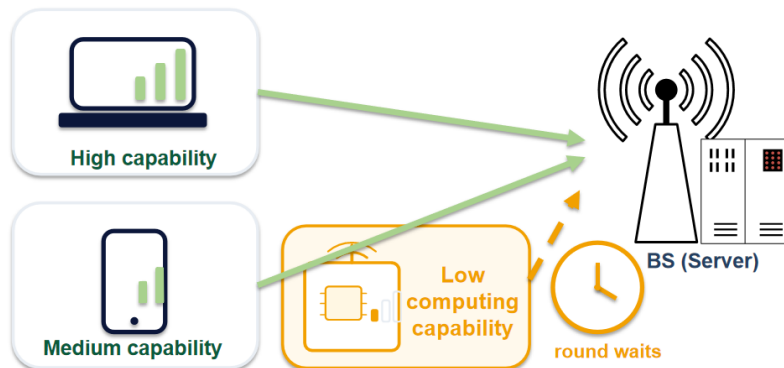
Federated Learning at the 6G Edge

6G envisions **collaborative edge intelligence** over ultra-dense and heterogeneous devices.

Federated learning (FL) enables *privacy-preserving* training: devices compute local updates and exchange only model information with the server.

Two trends drive the agenda:

- **scale**
- **heterogeneity**



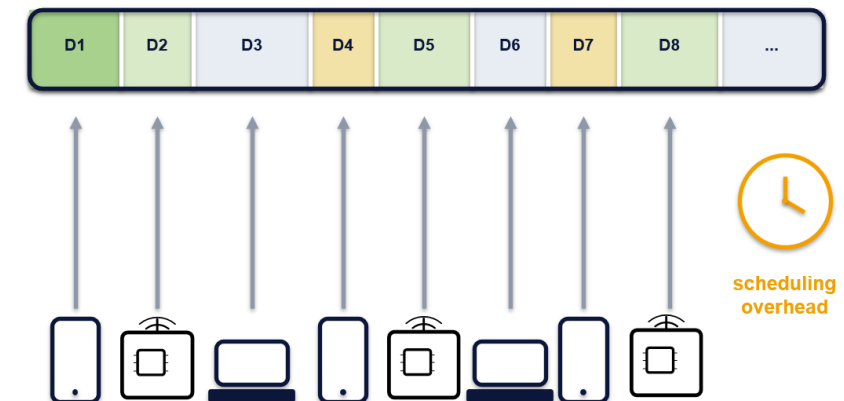
Two bottlenecks in conventional FL

1 Device heterogeneity causes straggler dominance

Resource-constrained devices *slow the entire round*, causing unstable training and degraded generalization.

2 OMA uplink becomes inefficient at scale

Fragmented bandwidth and scheduling overhead grow with the number of devices — a poor fit for *massive concurrent transmissions* envisioned in 6G.



Combining SFL and AirComp under Heterogeneity

Split Federated Learning (SFL)

Partition the DNN; offload deeper layers to the server.

→ Reduces on-device computation and *alleviates stragglers*.

Over-the-Air Computation (AirComp)

All devices transmit simultaneously, the channel sums signals via waveform superposition.

→ *Spectrum-efficient* and low-latency analog aggregation.

THE GAP

SFL with heterogeneous splits → **each layer has a different set of contributing devices.**

Standard AirComp-FL → **assumes all devices share the same sub-model.**

Our contributions

Heterogeneous AirComp-SFL

Per-device split points; AirComp-based layer-wise model aggregation.

MSE-constrained learning

Layer-wise MSE constraint bounds aggregation distortion, ensuring reliable convergence.

Joint optimization

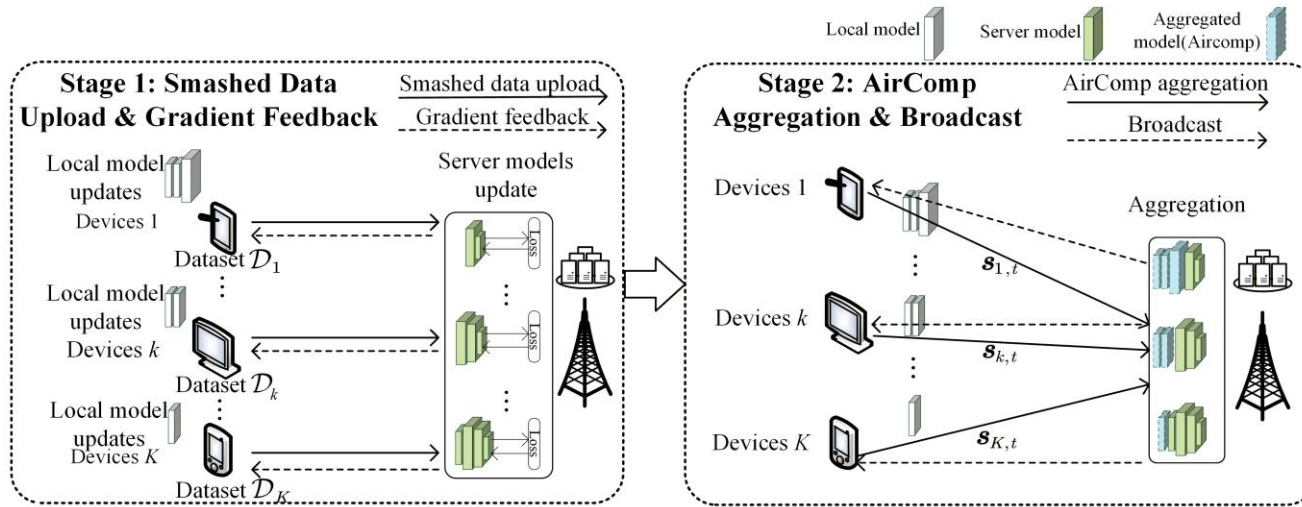
Latency minimization jointly across communication, computation, and split decisions.

Efficient algorithm

Decompose the coupled MINLP into tractable subproblems with closed-form and convex updates.

Joint Optimization Framework

System: two-stage SFL round



Stage 1: smashed-data upload (OMA) + server-side compute Stage 2: AirComp layer-wise aggregation

Problem: latency minimization

Decision variables:

Split points, bandwidth allocation, device and server computation resources, transmit power, and receive equalization.

Objective:

Minimize the per-round latency.

Subject to:

Reliable AirComp aggregation, device energy limits, and resource constraints.

Algorithm: alternating optimization

1. Resource allocation

Update bandwidth, CPU frequency, transmit power

Convex subproblem (Interior-point method)

2. Receive equalization

Update receive beamforming vector

Closed-form solution under Karush-Kuhn-Tucker (KKT) conditions

3. Split-point selection

Update device-specific split decisions

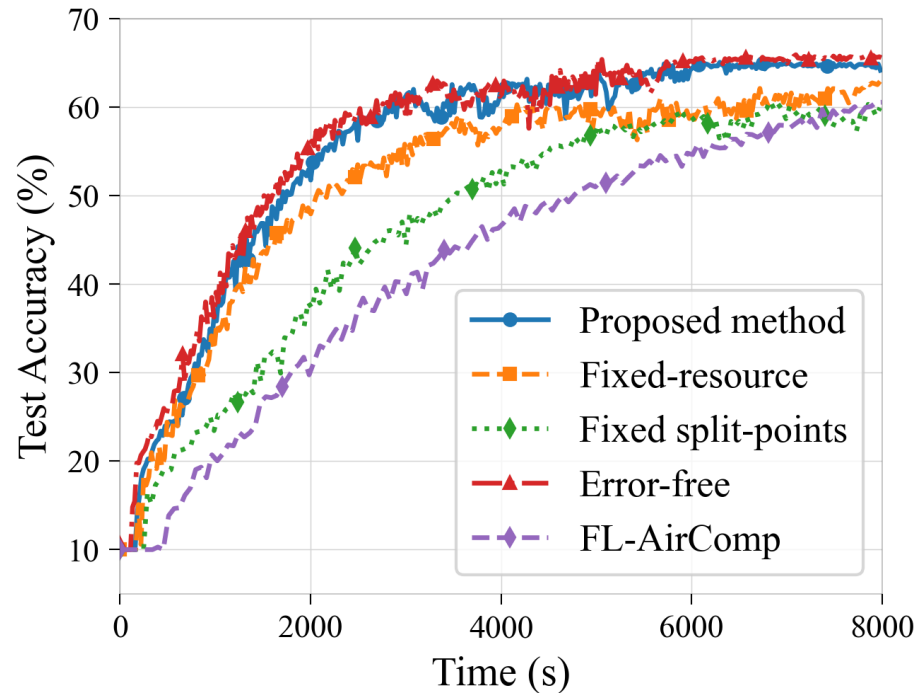
linear programming (LP) relaxation + rounding

Key idea: decompose the coupled MINLP into tractable subproblems.

Convergence and Robustness

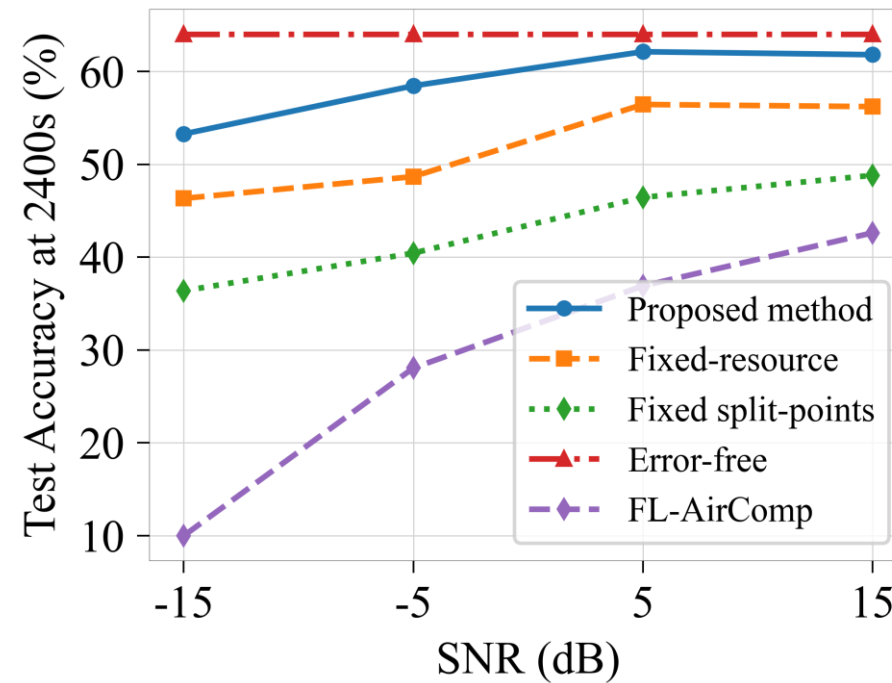
Setup K = 30 devices CIFAR-10 non-IID Rayleigh fading SNR = 5 dB

Baselines FL-AirComp Fixed-resource Fixed split-points Error-free



Faster convergence

The proposed method approaches the error-free upper bound and outperforms all baselines under non-IID data.



Robust to low SNR

The method remains stable when channel noise increases, while fixed-resource and FL-AirComp degrade more severely.

Thank you

Q & A

CONTACT

Haowen Sun

sunhaowen@seu.edu.cn

Southeast University
Trinity College Dublin

Full Result Set & Baselines

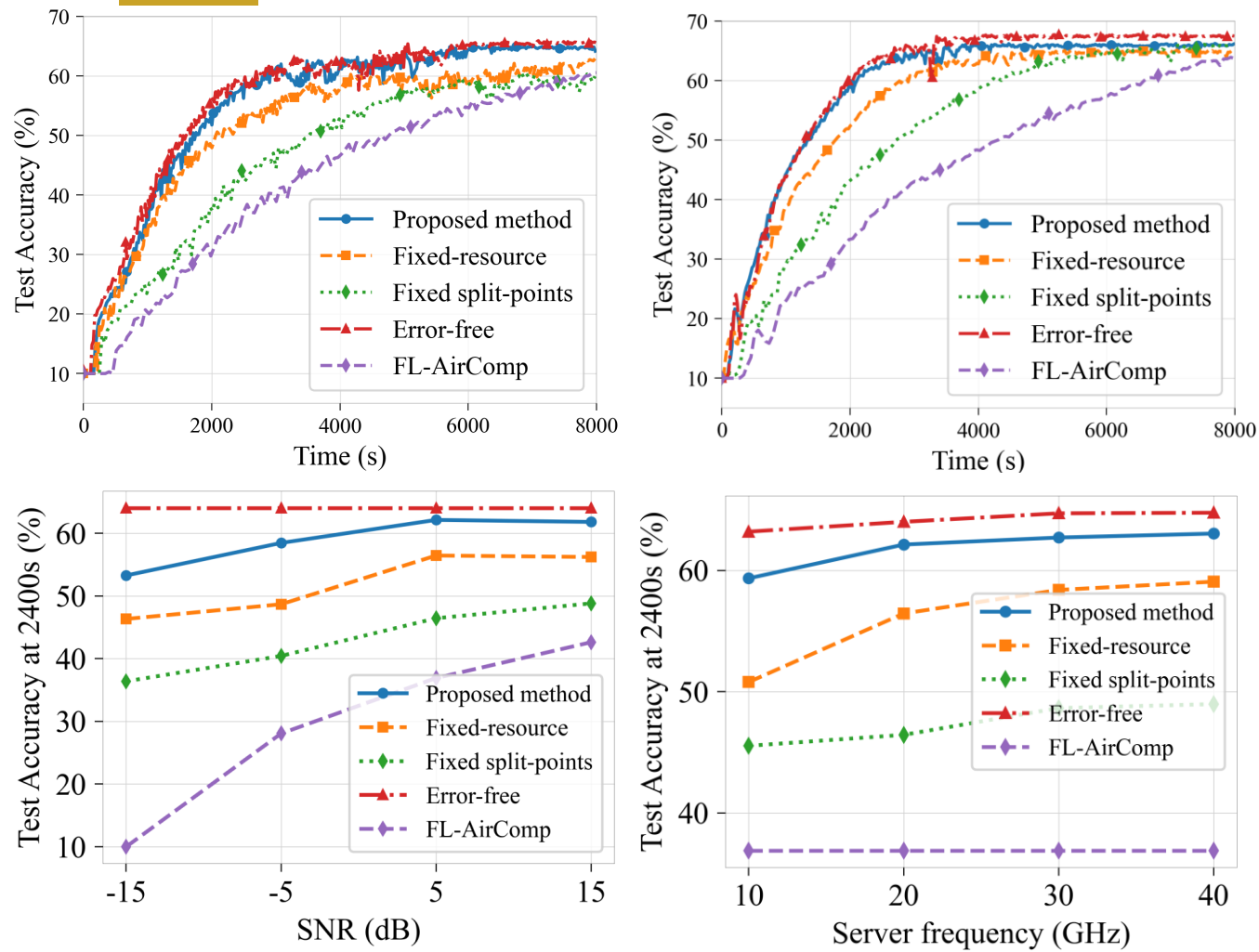


Fig. 2: (a) IID accuracy-time (b) non-IID accuracy-time (c) accuracy vs. SNR (d) accuracy vs. server CPU frequency.

Baselines

- Error-free** Noise-free aggregation; theoretical upper bound.
- FL-AirComp** AirComp aggregation without split or joint optimization.
- Fixed-resource** Optimize split points only; resources fixed.
- Fixed split-points** Optimize all other variables; split points fixed.

Observations

- Channel noise is the primary bottleneck; SNR drives accuracy gaps.
- FL-AirComp collapses near -15 dB; lacks split + joint optimization.
- Fixed-resource cannot tune power + Rx equalizer jointly $\rightarrow$ larger AirComp distortion.
- Fixed-split underuses both device and server compute, inflating latency.
- Proposed scheme tracks Error-free closely under both IID and non-IID data.