entropy

MDPI

Article

Tail-Latency-Aware Federated Learning with Pinching Antenna: Latency, Participation, and Placement

Yushen Lin [1,*] and Zhiguo Ding [2]

[1] Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK
[2] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore; zhiguo.ding@ntu.edu.sg
* Correspondence: yushen.lin@manchester.ac.uk

Abstract

Straggler synchronization is a dominant wall-clock bottleneck in synchronous wireless federated learning (FL). Under non-IID data, however, aggressively sampling only fast clients may significantly slow convergence due to statistical heterogeneity. This paper studies PASS-enabled FL, where a radiating pinching antenna (PA) can be activated at an arbitrary position along a dielectric waveguide to reshape uplink latencies. We consider a joint optimization of PA placement and client participation to minimize a proxy for time-to-accuracy, coupling the exact expected maximum round latency via order statistics with a heterogeneity-aware statistical-efficiency proxy. We derive first-order optimality conditions that reveal an explicit tail-latency premium in the KKT recursion, quantifying how latency gaps are amplified by maximum-order-statistic synchronization. Under a latency-class structure, we obtain a within-class square-root sampling law and establish a two-class phase transition where slow-class participation collapses under an explicit heterogeneity-threshold condition as the per-round sample size grows. For PA placement, we prove a piecewise envelope-derivative characterization and provide an exact breakpoint-and-root candidate-enumeration procedure. Simulation results validate the structural findings and show that PASS enables more eligible participation, yielding higher wall-clock accuracy.

Keywords: pinching antenna; federated learning (FL); non-IID; tail latency; stragglers; antenna placement

1. Introduction

Federated learning (FL) has attracted research interest due to its ability to address data privacy concerns while enabling collaborative learning over large-scale distributed data, as well as alleviating the heavy computational burden on end clients [1,2]. However, wireless FL is often wall-clock-limited by stragglers: in synchronous FL, each round must wait for the slowest selected uplink (the "straggler"), and the slowest-link tail dominates the overall training time [3]. Meanwhile, non-IID data create statistical heterogeneity, making the aggressive selection of "fast" devices potentially harmful to convergence [4].

Pinching-antenna systems (PASS) have recently gained significant research attention as a physical-layer mechanism to control the propagation environment [5]. PASS operates by dynamically activating radiating pinching antennas (PAs) along a dielectric waveguide. By activating PAs at different positions, PASS can create line-of-sight (LoS) links and establish short "last-meter" connections [6,7], which has the potential to mitigate stragglers by shortening the worst links and stabilizing LoS connectivity in challenging high-frequency environments [8,9].

Check for updates

Academic Editors: Shlomo Shamai (Shitz), Giuseppe Caire, Osvaldo Simeone, Holger Boche and Onur Günlü

Received: 25 January 2026
Revised: 9 March 2026
Accepted: 16 March 2026
Published: 18 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1.1. PASS

At a foundational level, ref. [6] studies pinching antennas within the broader class of flexible-antenna systems, developing analytical results for the basic single-waveguide/single-PA setting to quantify how position control can mitigate large-scale path loss by creating strong LoS links. Since PA location is the main control knob in PASS, placement optimization has also been studied to obtain sharper insights. In [10], closed-form placement solutions are derived for several multiple-access objectives. For user-fairness-oriented OMA, the optimal PA activation location must be beneficial to all served users and does not depend on users' perpendicular distances to the waveguide [10].

Joint resource allocation is challenging because PA locations couple with conventional variables such as transmit power and time/frequency resources [9]. The work in [11] proposes multi-waveguide transmission structures, including waveguide multiplexing, waveguide division, and waveguide switching. It formulates max–min fairness beamforming problems for multi-group multicast, and develops penalty dual decomposition-based algorithms to handle the complex exponential and fractional coupling induced by pinching beamforming [11]. In [12], a multi-waveguide PASS-assisted ISAC architecture is proposed, and sensing accuracy is measured by a CRB metric. The resulting CRB-minimization problem is solved by alternating between SDR-based digital beamforming and penalty-based PA deployment updates under communication QoS and deployment constraints [12].

1.2. Straggler and Dropout Mitigation in Federated Learning

In synchronous FL, the wall-clock time of each round is dictated by the slowest selected client, which makes training sensitive to stragglers. A large body of work mitigates this effect at the system and protocol layers, often without changing the underlying wireless channel. Client selection and scheduling are common approaches: Oort prioritizes clients using a utility-and-speed signal to improve training efficiency [13], while TiFL uses a tier-based design to reduce waiting time and to manage device heterogeneity [14]. Another direction relaxes strict synchronization. FedBuff introduces buffered asynchronous aggregation to reduce idle waiting and to better tolerate stragglers [15]. Recent wireless-edge work also uses straggler-aware grouping and clustering to stabilize training under heterogeneous link qualities [16]. From a learning-theoretic angle, straggler-resilient FL has been studied by explicitly balancing statistical accuracy with system heterogeneity, which clarifies when faster participation helps or hurts convergence [17]. These methods are effective in many settings, but they typically treat the channel and latency distribution as fixed. As a result, they often rely on biased participation (favoring fast clients) or asynchrony, both of which can interact poorly with non-IID data.

1.3. PASS-Empowered Edge Learning and PASS-FL

Unlike the above approaches, PASS can reshape the physical-layer latency landscape itself by controlling where radiation occurs. Different from reconfigurable intelligent surfaces (RISs) or movable antenna [18,19], PASS changes the effective radio "access point" location along the waveguide, thereby directly modifying the path-loss geometry and the identity of the straggler link [20]. This "geometry control" is particularly relevant for tail latency: since synchronous FL is governed by the maximum uplink latency, even a modest reduction in the worst-link distance (or elimination of NLoS blockage via a short last-meter LoS) can yield a disproportionate reduction in round time. Moreover, PASS can improve slow-client eligibility without relying solely on biased client selection or loosening synchronization, which is important under non-IID data. This capability has made PASS attractive for edge intelligence, where training time is dominated by extreme straggler rounds rather than average latency. A broader view on the synergy between pinching antennas and AI

argues that AI can control pinching positions, while pinching antennas can support edge-AI workloads such as FL and over-the-air aggregation [21,22]. At the aggregation level, PASS has been studied for AirComp to reduce channel misalignment and improve aggregation accuracy through joint placement and communication design [21]. For FL specifically, recent work analyzed PASS-enabled FL in terms of tail latency, participation probability, and convergence behavior, highlighting that addressing stragglers at the PHY layer can complement higher-layer scheduling [22]. Related studies also considered hybrid networks that combine conventional access and pinching antennas and combine optimized placement and resource allocation to improve communication efficiency for FL [23]. However, existing PASS-FL works focus primarily on demonstrating latency or participation improvements through system-level optimization; they do not provide analytical characterizations of the optimal participation–latency tradeoff under non-IID data or derive structural properties of the joint placement-and-sampling problem.

To position the present work, we briefly compare with existing straggler-aware FL and PASS-FL studies. Oort [13] relies on heuristic measurements without PHY-layer reshaping or joint optimization. TiFL [14] introduces tier-based waiting but offers only limited non-IID awareness and no joint structure. Reisizadeh et al. [17] model the system/learning tradeoff with full non-IID coupling, yet address sampling only (partial joint structure) and lack PHY reshaping. Lin et al. (PASS-FL) [22] first apply PASS to FL with emphasis on tail latency, but provide limited non-IID coupling and system-level optimization without analytical joint structure. In contrast, this work combines PHY latency reshaping via PASS with an exact order-statistics straggler model (Equations (6) and (7)), explicit non-IID coupling through $g(q)$, and a joint (q, x) characterization via KKT conditions and a placement envelope.

This paper addresses these gaps with the following contributions:

- We propose a cross-layer proxy objective for time-to-accuracy in PASS-enabled wireless FL under non-IID data by coupling exact straggler latency with a heterogeneity-aware statistical-efficiency proxy.
- For a fixed PA location, we characterize locally optimal client sampling via KKT conditions and reveal a tail-latency premium that quantifies straggler amplification across latency gaps.
- We derive a class-wise reduction that yields a within-class square-root law and establish a two-class phase transition where slow-class participation collapses to the $1/K$ scale under explicit conditions.
- We provide an exact piecewise envelope-derivative characterization for PA placement and develop a breakpoint-and-root candidate search that is globally optimal for the outer envelope objective.
- Simulations validate the phase transition, demonstrate Pareto-dominant operating points in the latency–statistical-efficiency trade-off, and show substantial accuracy gains under deadline-constrained FL.

1.4. Organization

Section 2 introduces the PASS link/latency model and the order-statistics straggler functional. Section 3 studies optimal client participation for a fixed PA location, derives the tail-latency-aware KKT conditions, and establishes the latency-class reduction and phase transition. Section 4 characterizes the PA placement problem via a piecewise envelope derivative and presents a breakpoint-and-root candidate search algorithm. Section 5 provides numerical experiments, followed by conclusions in Section 6.

2. System Model

We consider a single-waveguide PASS in which a radiating pinching antenna is activated at position x along the waveguide, where $L > 0$ denotes the length of the dielectric waveguide, so that the PA activation position satisfies $x \in [0, L]$. The single-antenna client i has the projection of the waveguide $u_i \in \mathbb{R}$ and the effective transverse distance $r_i > 0$ to the waveguide, where $i \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$ denotes the set of clients and N is the total number of clients. The PA–client distance is

$$d_i(x)^2 = (x - u_i)^2 + r_i^2. \tag{1}$$

Following the standard PASS link model [6,22], the end-to-end uplink channel gain consists of free-space LoS spherical-wave propagation from the PA to each client. Specifically, the complex baseband channel from the PA at x to client i is modeled as [21]

$$h_i(x) = \sqrt{\eta_f}\, \frac{e^{-jk_0 d_i(x)}}{d_i(x)}, \tag{2a}$$

$$\gamma_i(x) = \frac{P_i}{\sigma^2} |h_i(x)|^2, \tag{2b}$$

where P_i is the uplink transmit power of client i and σ^2 is the receiver noise power. Here, $\eta_f \triangleq \frac{c_0^2}{16\pi^2 f_c^2}$ is the Friis free-space factor, and $k_0 \triangleq 2\pi f_c / c_0$ is the free-space wavenumber.

Each selected client uploads S_i bits over bandwidth B. The per-round latency of client i is

$$t_i(x) = t_i^{\text{comp}} + \tau_i(x), \qquad \tau_i(x) = \frac{S_i}{B \log_2\big(1 + \gamma_i(x)\big)}, \tag{3}$$

where t_i^{comp} denotes the local computational time of client i.

For a fixed PA position x, we relabel clients so that their per-round latencies satisfy

$$t_1(x) \le t_2(x) \le \cdots \le t_N(x). \tag{4}$$

All quantities indexed by i refer to this sorted order. Through (2) and (3), PA placement x determines the uplink SNR $\gamma_i(x)$ via $d_i(x)$, and thus $t_i(x)$. Varying x reshapes $\{t_i(x)\}_{i=1}^N$ and reorders clients by latency, i.e., the identity of the slowest links can change with x. Since synchronous FL waits for the maximum latency among the K sampled clients, the distribution of the straggler depends on the relative ordering of $\{t_i(x)\}$ under the sampling probabilities q. This motivates the sorted relabeling in (4) and the order-statistics treatment below.

2.1. Straggler-Order Statistics

In each round, the server samples K clients i.i.d. with replacement according to a distribution $q = (q_1, \ldots, q_N)$, where

$$q \in \Delta \triangleq \Big\{ q \in \mathbb{R}^N : \; q_i > 0, \; \sum_{i=1}^{N} q_i = 1 \Big\}. \tag{5}$$

Let $\mathcal{K}(q)$ denote the multiset of sampled clients. The per-round wall-clock time is determined by the straggler: $T_{\text{round}}(q, x) = \max_{i \in \mathcal{K}(q)} t_i(x)$. Define the cumulative sampling mass up to client i in the sorted order in (4):

$$Q_i \triangleq \sum_{j=1}^{i} q_j, \qquad Q_0 \triangleq 0. \tag{6}$$

Then, it is straightforward to obtain that, for fixed x and sorted times as in (4),

$$f(q,x) \triangleq \mathbb{E}\big[T_{\text{round}}(q,x)\big] = \sum_{i=1}^{N} \big(Q_i^K - Q_{i-1}^K\big)\, t_i(x). \tag{7}$$

2.1.1. Tail-Latency Coupling

Before analyzing the general N-client problem, we illustrate the mechanism of tail-latency amplification using a two-class model. Consider a fast class with latency t_f and a slow class with latency $t_s = t_f + \Delta$. Let $\delta \in [0,1]$ be the total sampling mass assigned to the slow class. Under i.i.d. sampling,

$$\mathbb{E}[T_{\text{round}}] = t_f + \Delta\big(1-(1-\delta)^K\big), \qquad \frac{\partial}{\partial\delta}\mathbb{E}[T_{\text{round}}] = \Delta K(1-\delta)^{K-1}.$$

When $\delta K \ll 1$, we have $(1-\delta)^{K-1} = 1 + O(\delta K)$ and thus $\frac{\partial}{\partial\delta}\mathbb{E}[T_{\text{round}}] = \Delta K\,(1+O(\delta K))$.

The term $1-(1-\delta)^K$ is the probability that at least one draw lands in the slow class. The factor $(1-\delta)^{K-1}$ is the probability that the other $K-1$ draws remain in the fast class, so the marginal cost $\frac{\partial}{\partial\delta}\mathbb{E}[T_{\text{round}}] = \Delta K(1-\delta)^{K-1}$ can be read as *(gap height)* $\times$ *(K-fold amplification)* $\times$ *("near-miss" probability)*. This also reveals the natural participation scale $\delta = \rho/K$; then, $(1-\delta)^K \to e^{-\rho}$ and the slow class ceases to be an almost-sure straggler.

This two-class intuition of K-fold amplification of latency gaps extends to the general N-client setting and is formalized in Section 3.

2.1.2. Heterogeneity-Aware Statistical Proxy

Consider the global FL objective

$$F(\boldsymbol{w}) = \sum_{i=1}^{N} p_i F_i(\boldsymbol{w}), \qquad \sum_{i=1}^{N} p_i = 1,\ p_i > 0, \tag{8}$$

where p_i denotes the target aggregation weight, e.g., $p_i = n_i / \sum_j n_j$, and client i holds a local dataset $\mathcal{D}_i = \{\xi_{i,1}, \ldots, \xi_{i,n_i}\}$ of size n_i. The local empirical risk is

$$F_i(\boldsymbol{w}) \triangleq \frac{1}{n_i}\sum_{k=1}^{n_i} \ell(\boldsymbol{w};\xi_{i,k}), \qquad i \in \{1,\ldots,N\}. \tag{9}$$

where $\ell(\boldsymbol{w};\xi)$ denotes the per-sample loss function. The goal of FL is to solve

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} F(\boldsymbol{w}). \tag{10}$$

We denote by $\nabla F_i(\boldsymbol{w};\xi_i)$ the stochastic gradient computed from a random sample ξ_i (or a mini-batch) drawn from $\mathcal{D}_i$, i.e., $\nabla F_i(\boldsymbol{w};\xi_i) = \nabla\ell(\boldsymbol{w};\xi_i)$ and $\mathbb{E}_{\xi_i}[\nabla F_i(\boldsymbol{w};\xi_i)] = \nabla F_i(\boldsymbol{w})$. In each round r, the server samples K clients i.i.d. according to $\boldsymbol{q} = (q_1,\ldots,q_N) \in \Delta$ and broadcasts the current global model $\boldsymbol{w}^{(r)}$ under FedAvg. Each selected client i performs E local SGD steps:

$$\boldsymbol{w}_i^{(r,0)} = \boldsymbol{w}^{(r)}, \tag{11}$$

$$\boldsymbol{w}_i^{(r,e+1)} = \boldsymbol{w}_i^{(r,e)} - \eta\,\nabla F_i(\boldsymbol{w}_i^{(r,e)};\xi_i^{(r,e)}), \qquad e = 0,\ldots,E-1, \tag{12}$$

and returns $\boldsymbol{w}_i^{(r,E)}$.

To capture the statistical cost of biased participation under non-IID data, we adopt the sampling-dependent heterogeneity proxy

$$g(q) \triangleq \omega \sum_{i=1}^{N} \frac{c_i}{q_i} + \nu, \tag{13}$$

where $c_i > 0$ captures the relative statistical importance of client i and $\omega, \nu > 0$ are design constants. The $1/q_i$ dependence is motivated by standard FL analyses under simplifying assumptions [24–26], but in the present paper it is used only as a design surrogate rather than as a literal convergence guarantee for the non-convex CNN experiments in Section 5. Accordingly, we define the cross-layer objective

$$J(q, x) \triangleq f(q, x)\, g(q), \tag{14}$$

and interpret $J(q, x)$ as a proxy for time-to-accuracy. The PASS latency model, the exact order-statistics straggler functional $f(q, x)$, the tail-sensitivity/KKT structure, the phase-transition analysis, and the PA-placement characterization require only the explicit form of $g(q)$; they do not rely on the convexity of the experimental learning model.

3. Optimal Participation for PA Location

Efficient wireless FL requires balancing per-round latency against statistical efficiency: minimizing the duration of each round while controlling the total number of rounds suggested by the heterogeneity proxy. In this section, we optimize the client sampling distribution for a fixed PA position x. To simplify notation, we suppress the dependence on x and write $t_i \triangleq t_i(x)$, $f(q) \triangleq f(q, x)$ and $J(q) \triangleq J(q, x)$ under the sorted order in (4). We study the inner problem

$$(\mathbf{P1}) \qquad \min_{q \in \Delta} J(q) \triangleq f(q) g(q). \tag{15}$$

Problem (P1) is generally nonconvex due to the product coupling of a concave straggler functional $f(\cdot)$ and a convex heterogeneity proxy term $g(\cdot)$. A first question is whether an optimizer of (P1) could lie on the simplex boundary, i.e., assign $q_i = 0$ to some clients, or whether the non-IID term forces every q_i to stay strictly positive.

Lemma 1. *Suppose $c_i > 0$ for all i. Then, $J(q) = f(q)g(q)$ satisfies $\lim_{q_i \to 0^+} J(q) = +\infty$ for any i. Consequently, problem (P1) admits at least one global minimizer $q^{\mathrm{opt}} \in \Delta$, and every global minimizer is bounded away from the boundary; in particular, it is an interior point.*

Proof. Refer to Appendix A. □

The next question is whether the exact order-statistics form can be rewritten into a representation that separates the contribution of each adjacent latency gap and makes the dependence on $\{Q_i\}$ explicit. To solve the inner minimization over q for a fixed x, we rewrite $f(q)$ in a telescoping form over adjacent latency gaps, which makes its dependence on q explicit. Starting from (7), expand $\sum_{i=1}^{N} (Q_i^K - Q_{i-1}^K) t_i = t_N Q_N^K + \sum_{i=1}^{N-1} Q_i^K (t_i - t_{i+1})$:

$$f(q) = t_N - \sum_{i=1}^{N-1} \Delta_i\, Q_i^K. \tag{16}$$

where $\Delta_i \triangleq t_{i+1} - t_i \geq 0$ and $Q_N = 1$. In (16), Q_i^K is the probability that all K i.i.d. draws fall within the prefix $\{1, \ldots, i\}$; thus, each gap Δ_i reduces the baseline t_N only when the sampled maximum does not cross that latency cliff (an event of probability Q_i^K). Equation (16) shows that straggler latency is governed by a sum of latency "cliffs" Δ_i, each discounted by the

probability Q_i^K that all K samples fall within the faster prefix. This makes it explicit which gaps matter (those near the tail with $Q_i \approx 1$). This raises a natural question: what is the marginal effect of increasing q_s on $f(q)$, i.e., how sensitive is the tail latency to each coordinate of q? The following lemma can be obtained.

Lemma 2. *Fix x and consider the sorted latencies $t_1 \le \cdots \le t_N$. For $q \in \bar{\Delta} \triangleq \{q \in \mathbb{R}^N : q_i \ge 0, \sum_{i=1}^N q_i = 1\}$, define $Q_i \triangleq \sum_{j=1}^i q_j$ and $\Delta_i \triangleq t_{i+1} - t_i \ge 0$. Then, the expected straggler latency admits the gap form in (16). Moreover, the following hold: (i) The mapping $q \mapsto f(q)$ is concave on $\bar{\Delta}$ (and hence on Δ). (ii) For any interior $q \in \Delta$, the directional derivative of f along any feasible direction v (where $\sum_{i=1}^N v_i = 0$) is given by*

$$\mathrm{D}f(q)[v] = -\sum_{s=1}^{N} D_s(q)\, v_s, \qquad D_s(q) \triangleq K \sum_{i=s}^{N-1} \Delta_i Q_i^{K-1} \ (\ge 0). \tag{17}$$

This provides a convenient gradient representative on the simplex, such that one may take $\frac{\partial f}{\partial q_s} = -D_s(q)$; (iii) For $K \ge 2$, f is twice continuously differentiable on Δ with Hessian

$$\nabla^2 f(q) = -K(K-1) \sum_{i=1}^{N-1} \Delta_i Q_i^{K-2}\, a_i a_i^\top \preceq 0, \qquad a_i \triangleq (\underbrace{1, \ldots, 1}_{i}, 0, \ldots, 0)^\top. \tag{18}$$

Proof. Refer to Appendix B. □

Using this gradient structure, we derive the KKT system, whose recursion explicitly shows how the optimal q trades off statistical weights c_i against the marginal tail penalty induced by straggling. We now ask, compared with the statistics-only square-root rule $q_i \propto \sqrt{c_i}$, how is the optimal sampling distribution distorted by straggler synchronization, and which latency gaps Δ_i govern this distortion. From (16), any feasible perturbation of q satisfies $\sum_i \mathrm{d}q_i = 0$ and hence $\mathrm{d}Q_N = 0$, so the directional derivative of f along the simplex equals the derivative of the reduced form (16): $\partial f / \partial q_s = -D_s(q)$. Also, $\frac{\partial g}{\partial q_s} = -\omega \frac{c_s}{q_s^2}$. The Lagrangian $\mathcal{L}(q, \mu) = f(q)g(q) + \mu(\sum_i q_i - 1)$ yields the stationarity condition $g\, \partial f / \partial q_s + f\, \partial g / \partial q_s + \mu = 0$. Substituting derivatives gives (19) with $\lambda \triangleq \mu$. For (20), subtract (19) for $s = i$ and $s = i + 1$ and use $D_i(q) - D_{i+1}(q) = K\Delta_i Q_i^{K-1}$.

Proposition 1. *Any interior KKT point $q^\star \in \Delta$ of (P1) satisfies, for some scalar λ,*

$$f(q^\star)\, \omega\, \frac{c_s}{(q_s^\star)^2} + g(q^\star) D_s(q^\star) = \lambda, \qquad \forall s \in \mathcal{N}. \tag{19}$$

where $D_s(q) \triangleq K \sum_{i=s}^{N-1} \Delta_i\, Q_i^{K-1} \ge 0$ denotes tail sensitivity. Moreover, adjacent indices satisfy

$$\frac{c_{i+1}}{(q_{i+1}^\star)^2} = \frac{c_i}{(q_i^\star)^2} + \frac{g(q^\star)}{f(q^\star)\omega}\, K\, \Delta_i\, Q_i (q^\star)^{K-1}, \qquad i = 1, \ldots, N-1. \tag{20}$$

Remark 1. *If one minimizes the statistical factor $g(q)$ alone, the KKT conditions equalize c_i / q_i^2 across active clients, yielding the classical square-root rule $q_i \propto \sqrt{c_i}$. Under the coupled proxy objective $J(q) = f(q)g(q)$, moving sampling mass from a faster client i to a strictly slower client $i + 1$ must overcome an additional positive term:*

$$\frac{c_{i+1}}{(q_{i+1}^\star)^2} = \frac{c_i}{(q_i^\star)^2} + \underbrace{\frac{g(q^\star)}{f(q^\star)\omega} K \Delta_i Q_i (q^\star)^{K-1}}_{\text{tail-latency premium}}, \qquad i = 1, \ldots, N-1. \tag{21}$$

Here, $\Delta_i = t_{i+1} - t_i$ is the latency gap. The factor $Q_i(q^\star)^{K-1}$ admits a direct probabilistic interpretation: it is the probability that the other $K-1$ i.i.d. draws lie in the prefix $\{1,\dots,i\}$, so that the remaining draw becomes the decisive one that crosses the gap Δ_i and determines the round maximum. The multiplicative factor K reflects the K exchangeable opportunities for such a crossing.

This tail-latency premium is negligible unless Q_i is close to one. In particular, the effective amplification scales as KQ_i^{K-1}, which remains order-one only when $Q_i = 1 - \Theta(1/K)$, foreshadowing the $\Theta(1/K)$ participation scale revealed by the two-class phase transition in Theorem 1.

Let $q^\star \in \Delta$ be any interior KKT point in Proposition 1. Define $\psi_i \triangleq c_i/(q_i^\star)^2$. Then, $\{\psi_i\}_{i=1}^N$ is nondecreasing given Recursion (20):

$$\psi_{i+1} - \psi_i = \frac{g(q^\star)}{f(q^\star)\omega} K \Delta_i Q_i(q^\star)^{K-1} \geq 0, \qquad i = 1,\dots,N-1, \tag{22}$$

with strict inequality whenever $\Delta_i > 0$. Equivalently, $q_i^\star/\sqrt{c_i}$ is nonincreasing in i.

Remark 2. *Minimizing $g(q)$ alone yields the square-root rule $q_i \propto \sqrt{c_i}$. Under the coupled objective, Recursion (20) implies that $\psi_i \triangleq c_i/(q_i^\star)^2$ is nondecreasing with the latency index i; equivalently, $q_i^\star/\sqrt{c_i}$ is nonincreasing. Hence, slower clients are progressively reduced in relation to statistics-only allocation, with the distortion controlled by $K\Delta_i Q_i^{K-1}$.*

Class-Wise Reduction

When many clients share identical $t_i(x)$, we group them into latency classes, which enables a low-dimensional reformulation and yields a closed-form within-class allocation. The key question is whether, within a latency class where $\Delta_i = 0$, the KKT recursion simplifies to a tractable rule so that the N-dimensional vector q can be reduced to class-level masses.

Lemma 3. *Consider a fixed x; the sorted times take M distinct values $t^{(1)} < \cdots < t^{(M)}$, inducing a partition $\mathcal{N} = \bigcup_{m=1}^M \mathcal{C}_m$ where $\mathcal{C}_m = \{i : t_i = t^{(m)}\}$. Then, for each class m, any KKT point $q^\star$ satisfies*

$$q_i^\star = \delta_m^\star \frac{\sqrt{c_i}}{\sum_{j\in\mathcal{C}_m}\sqrt{c_j}}, \qquad \forall i \in \mathcal{C}_m, \tag{23}$$

where $\delta_m^\star \triangleq \sum_{i\in\mathcal{C}_m} q_i^\star$ is the total probability mass assigned to class m.

Proof. If i and $i+1$ belong to the same class, then $\Delta_i = 0$. Recursion (20) reduces to $c_{i+1}/(q_{i+1}^\star)^2 = c_i/(q_i^\star)^2$, implying $q_i^\star/\sqrt{c_i}$ is constant within the class. Normalization yields (23). □

Under Lemma 3, we can obtain the following corollary.

Corollary 1. *(P1) is equivalent to the M-dimensional problem*

$$\min_{\delta_1,\dots,\delta_M>0:\ \sum_m \delta_m = 1} \left(\sum_{m=1}^M \left(P_m^K - P_{m-1}^K\right) t^{(m)}\right)\left(\omega \sum_{m=1}^M \frac{C_m^2}{\delta_m} + \nu\right), \tag{24}$$

where $P_m \triangleq \sum_{\ell=1}^m \delta_\ell$ and $P_0 \triangleq 0$, $C_m \triangleq \sum_{i\in\mathcal{C}_m}\sqrt{c_i}$, $\delta_m \triangleq \sum_{i\in\mathcal{C}_m} q_i$, $\sum_{m=1}^M \delta_m = 1$.

Proof. By Lemma 3, $\sum_i c_i/q_i = \sum_m C_m^2/\delta_m$. The probability that the sampled maximum equals $t^{(m)}$ is $P_m^K - P_{m-1}^K$, giving the first factor. □

Having characterized $q^\star(x)$, we treat $J^\star(x) = J(q^\star(x), x)$ as piecewise smooth under a fixed latency ordering and derive its envelope derivative to obtain a first-order placement condition. Before turning to the outer placement problem in Section 4, we further specialize the reduced form in (24) to a two-class structure to expose an explicit participation scale and a phase transition in slow-class sampling.

Definition 1. *Let $\mathcal{F}$ and $\mathcal{S}$ be the fast/slow classes with times $t_f < t_s$ and gap $\Delta \triangleq t_s - t_f > 0$. Define*

$$C_f \triangleq \sum_{i\in\mathcal{F}} \sqrt{c_i}, \qquad C_s \triangleq \sum_{i\in\mathcal{S}} \sqrt{c_i}, \qquad \delta \triangleq \sum_{i\in\mathcal{S}} q_i \in (0,1). \tag{25}$$

Throughout this subsection, we use δ without a subscript for the simplicity of notation to denote the total probability mass assigned to the slow class, i.e., $\delta \triangleq \sum_{i\in\mathcal{S}} q_i$.

Lemma 3 fixes the within-class form in (26) for any given δ. Substituting into f yields $f(\delta) = t_f(1-\delta)^K + t_s\big(1-(1-\delta)^K\big) = t_s - \Delta(1-\delta)^K$. Similarly, $\sum_i c_i/q_i = C_f^2/(1-\delta) + C_s^2/\delta$, giving (27). Differentiation yields (28).

Under the two-class model, the optimal sampling distribution has the form

$$q_i^\star = \begin{cases} \dfrac{1-\delta^\star}{C_f}\sqrt{c_i}, & i \in \mathcal{F}, \\ \dfrac{\delta^\star}{C_s}\sqrt{c_i}, & i \in \mathcal{S}, \end{cases} \tag{26}$$

where $\delta^\star \in (0,1)$ minimizes the scalar objective

$$J(\delta) = \underbrace{\big[t_s - \Delta(1-\delta)^K\big]}_{f(\delta)} \underbrace{\left[\omega\Big(\frac{C_f^2}{1-\delta} + \frac{C_s^2}{\delta}\Big) + \nu\right]}_{g(\delta)}. \tag{27}$$

Any interior stationary point satisfies

$$\Delta K(1-\delta)^{K-1} g(\delta) + f(\delta)\,\omega\Big(\frac{C_f^2}{(1-\delta)^2} - \frac{C_s^2}{\delta^2}\Big) = 0. \tag{28}$$

Proposition 2. *Let $q^\star$ be any interior KKT point of the two-class problem. Then, $\delta^\star$ satisfies the scalar condition*

$$\frac{C_s^2}{(\delta^\star)^2} - \frac{C_f^2}{(1-\delta^\star)^2} = \frac{g(\delta^\star)}{\omega f(\delta^\star)}\, K\Delta\,(1-\delta^\star)^{K-1}, \tag{29}$$

with $f(\delta)$ and $g(\delta)$ defined in (27). In particular, the right-hand side is nonnegative; hence, any KKT point satisfies $C_s/(\delta^\star) \geq C_f/(1-\delta^\star)$.

Proof. Refer to Appendix C. □

Finally, we ask how the optimal slow-class mass behaves as the per-round sample size K grows: does straggler synchronization force $\delta^\star$ to shrink to the critical scale $\Theta(1/K)$, and under what explicit condition does such a collapse occur for global minimizers?

We obtain the following theorem, which is stated for a sequence of problems indexed by K. Since $C_s = \sum_{i\in\mathcal{S}} \sqrt{c_i}$ aggregates the slow-class statistical weight, the condition $C_s^2 = O(1/K)$ should be read as a small-slow-class regime in which the slow class contributes vanishing aggregated statistical benefit as K grows, e.g., the slow clients carry a diminishing share of the target weights p_i, so that the K-fold tail-latency amplification can dominate and force $\delta^\star$ to the critical $1/K$ scale.

Theorem 1. *Consider a sequence of two-class problems as in (27) indexed by the per-round sample size $K \to \infty$. Fix a constant $\rho > 0$ and define $\delta_\rho \triangleq \rho/K$ and $P_{\rho,K} \triangleq (1-\rho/K)^K$. Assume the slow-class aggregate satisfies the scaling*

$$C_s^2 = O(K^{-1}), \tag{30}$$

and, for some margin $\xi \in (0,1)$ and all sufficiently large K,

$$C_s^2 \le (1-\xi)\,\frac{\rho}{K}\,\frac{\Delta P_{\rho,K}}{t_s - \Delta P_{\rho,K}}\,\frac{\omega C_f^2 + \nu}{\omega}. \tag{31}$$

Then, any sequence of global minimizers $\{\delta_K^\star\}$ of (27) satisfies $\sup_K K\delta_K^\star < \infty$, i.e., the slow-class total probability mass collapses from $\Theta(1)$ to $\delta_K^\star = O(1/K)$. Moreover, Threshold (31) is explicit in $(\{c_i\}, \Delta, K, \omega, \nu)$ through (C_f, C_s).

Proof. Refer to Appendix D. □

Theorem 1 formalizes a participation collapse phenomenon driven by synchronous order statistics. For a fixed finite network, the theorem provides a sufficient condition for collapse at moderate K. When K grows, any constant slow-class mass $\delta = \Theta(1)$ makes the event "at least one slow client is selected" almost certain, so the expected round latency saturates at the slow-class time t_s. To keep the expected straggler latency strictly below t_s, the slow-class mass must shrink to the critical scale $\delta = \Theta(1/K)$, for which $(1-\delta)^K$ remains order one.

Condition (31) is a sufficient threshold on the slow-class aggregate heterogeneity C_s^2 relative to the fast class and system parameters. Intuitively, if the slow class contributes relatively little to the non-IID penalty, i.e., small C_s, then its statistical benefit cannot compensate the K-amplified tail latency it induces, and the wall-clock optimal policy samples it only at rate $O(1/K)$.

Remark 3. *The phase transition is driven by order statistics. Under two classes, the slow class becomes the straggler with probability $1-(1-\delta)^K$. If $\delta = \Theta(1)$, then $(1-\delta)^K \to 0$ as $K \to \infty$ and $f(\delta) = t_s - \Delta(1-\delta)^K \to t_s$, i.e., the slow class is an almost-sure straggler and the wall-clock latency saturates at the slow-class round time. To keep $f(\delta)$ strictly below t_s as K grows, the slow-class mass must shrink as $\delta = \rho/K$, for which $(1-\delta)^K \to e^{-\rho}$ and the slow class stops being an almost-sure maximum. Theorem 1 formalizes this intuition by providing an explicit sufficient condition under which any global minimizer satisfies $\delta^\star = O(1/K)$.*

Remark 4. *The wall-clock optimal design may drive a slow class to $\delta^\star = \Theta(1/K)$, which is often indistinguishable from excluding the class in practice. If slow clients correspond to long-tail data patterns needed for robust generalization, one can enforce coverage by adding a mild smoothing/coverage term (e.g., an entropy or KL regularizer) or by imposing lower-bound constraints such as $q_i \ge \eta p_i$ for some $\eta \in (0,1)$. Such modifications preserve the same order-statistics latency structure in $f(\cdot)$ and alter the KKT system by adding an additional marginal term that counteracts collapse, yielding an explicit knob to trade off wall-clock speed against distributional coverage.*

4. Pinching-Antenna Placement: Exact Characterization and Global Search

Section 3 characterized the optimal sampling distribution $q^*(x)$ for any fixed antenna position x. We now address the outer optimization: finding the globally optimal placement x^* that minimizes the resulting proxy objective $J^*(x) \triangleq \min_{q\in\Delta} J(q,x)$.

The primary challenge arises from the non-smoothness of $J^*(x)$, due to discrete changes in the latency ordering in (4) as x varies. Our strategy is to partition $[0, L]$ into finitely many regions (Lemma 4) within which the ordering is fixed, apply an envelope-theorem derivative (Proposition 3), and search over stationary points and region boundaries (Algorithm 1).

Algorithm 1 Tail-latency PASS placement and participation.

Require: $\{u_i, r_i, S_i, t_i^{\text{comp}}, c_i\}_{i=1}^N$, bandwidth B, ω, ν, K, interval $[0, L]$, tolerance η.
Ensure: x^* and sampling distribution $q^\star$.
1: Compute $t_i(x)$ by (3) and $t_i'(x)$ by (32).
2: Initialize breakpoint set $\mathcal{B} \leftarrow \{0, L\}$.
3: **for** each pair (i, j) with $i < j$ **do**
4: Numerically solve $t_i(x) = t_j(x)$ on $[0, L]$; add all roots to $\mathcal{B}$ (within tolerance η).
5: **end for**
6: Sort $\mathcal{B}$ and form intervals $\mathcal{I}_m = [\mathcal{B}_m, \mathcal{B}_{m+1}]$.
7: Initialize best value $J_{\min} \leftarrow \infty$ and best solution $(x^\star, q^\star) \leftarrow (\text{null}, \text{null})$.
8: **for** each interval $\mathcal{I}_m$ **do**
9: Pick $x_0 \in \text{int}(\mathcal{I}_m)$ and determine the fixed ordering of $\{t_i(x_0)\}$.
10: **Inner solve:** for each queried x, compute minimizer $\hat{q}(x)$ of $\min_{q \in \Delta} J(q, x)$:
11: (a) if multiple clients have identical latencies, form latency classes and solve the reduced problem (24);
12: (b) otherwise, solve **(P1)** using a multi-start strategy (e.g., projected gradient/SQP initialized from the statistical rule $q_i \propto \sqrt{c_i}$ and from perturbed variants), and keep the best solution found.
13: **Outer root:** define $\phi(x) \triangleq \sum_i (Q_i(\hat{q}(x))^K - Q_{i-1}(\hat{q}(x))^K)\, t_i'(x)$.
14: Find all roots of $\phi(x) = 0$ in $\mathcal{I}_m$ (e.g., bracketing + bisection) and collect candidates $\mathcal{X}_m$.
15: Evaluate $J(\hat{q}(x), x)$ for all $x \in \mathcal{X}_m \cup \{\mathcal{B}_m, \mathcal{B}_{m+1}\}$ and update $(x^\star, q^\star)$ if improved.
16: **end for**
17: **return** $x^\star, q^\star$.

4.1. Envelope Derivative

The PA position affects only $f(q, x)$ through the times $t_i(x)$. Ordering changes (i.e., $t_i(x) = t_j(x)$) create nondifferentiable points because the sorted order in (4) changes. Between such breakpoints, the ordering is fixed. For the placement derivative we need $t_i'(x)$. From (3), we have

$$t_i'(x) = \frac{S_i}{B \ln 2} \frac{\gamma_i(x)}{(1 + \gamma_i(x)) \left[\log_2(1 + \gamma_i(x))\right]^2} \left(\frac{2(x - u_i)}{(x - u_i)^2 + r_i^2} \right). \tag{32}$$

Proposition 3. *Fix an interval $\mathcal{I} \subset [0, L]$ on which the ordering in (4) does not change. Let $q^\star(x) \in \arg\min_{q \in \Delta} J(q, x)$. Assume $q^\star(x)$ is chosen so that it varies continuously on $\mathcal{I}$ (e.g., the inner minimizer is unique). Then, for all $x \in \mathcal{I}$ where $t_i(x)$ are differentiable,*

$$\frac{\mathrm{d}}{\mathrm{d}x} J^\star(x) = g(q^\star(x)) \sum_{i=1}^N \left(Q_i(q^\star(x))^K - Q_{i-1}(q^\star(x))^K\right) t_i'(x). \tag{33}$$

Consequently, any stationary point $x^\star \in \text{int}(\mathcal{I})$ satisfies

$$\sum_{i=1}^N \left(Q_i(q^\star(x^\star))^K - Q_{i-1}(q^\star(x^\star))^K\right) t_i'(x^\star) = 0, \tag{34}$$

with $t_i'(x)$ given by (32).

Proof. On $\mathcal{I}$, the ordering in (4) is fixed, so the straggler expectation admits the smooth representation $f(q,x) = \sum_{i=1}^{N} \pi_i(q)\, t_i(x)$ with weights $\pi_i(q) \triangleq Q_i^K - Q_{i-1}^K$ that depend only on q. Therefore, $\partial J/\partial x = g(q) \sum_i \pi_i(q)\, t_i'(x)$.

Although the inner minimization over q is generally nonconvex due to the product coupling, $J(q,x)$ is continuously differentiable in x on $\mathcal{I}$. Under the standing assumption that a continuous selection $q^\star(x) \in \arg\min_{q\in\Delta} J(q,x)$ exists on $\mathcal{I}$ (e.g., when the inner minimizer is unique), a standard envelope-theorem argument yields $\frac{\mathrm{d}}{\mathrm{d}x} J^\star(x) = \frac{\partial J(q,x)}{\partial x}\Big|_{q=q^\star(x)}$, which gives (33). Setting (33) to zero yields the stationary condition in (34). □

Remark 5. *The weight $\pi_i(q^\star(x)) = Q_i(q^\star(x))^K - Q_{i-1}(q^\star(x))^K$ equals the probability that client i is the straggler (the round maximum) under sampling distribution $q^\star(x)$. Thus, Equation (33) states that the outer gradient is a weighted sum of physical gradients $t_i'(x)$, dominated by clients that are most likely to determine the synchronous round time.*

We justify this piecewise analysis by showing that order changes occur only at finitely many pairwise crossings of $t_i(x)$, yielding a finite breakpoint partition of $[0, L]$. Without loss of generality, we assume that, for any pair $i \neq j$, the function $t_i(x) - t_j(x)$ is not identically zero on $[0, L]$.

Lemma 4. *Under the considered PASS model, for each pair (i, j), the equation $t_i(x) = t_j(x)$ has finitely many solutions in $[0, L]$. Hence, the set of ordering breakpoints is finite.*

Proof. The functions $t_i(x)$ are real-analytic in x on $[0, L]$ because they are compositions of analytic functions (polynomials, exponential, and $\log(1 + \cdot)$). Thus, $t_i(x) - t_j(x)$ is analytic; if it is not identically zero, it has only finitely many zeros on the compact interval $[0, L]$. □

4.2. Global Placement Algorithm

In this breakpoint partition, we evaluate stationary points and interval boundaries to assemble a breakpoint-and-root candidate-enumeration algorithm for PA placement. By Lemma 4, $[0, L]$ can be partitioned into finitely many fixed-ordering intervals. In each interval, Proposition 3 reduces placement to 1D root finding.

Algorithm 1 is globally optimal for the outer envelope problem $\min_{x\in[0,L]} J^\star(x)$ with $J^\star(x) = \min_{q\in\Delta} J(q,x)$, provided that the inner participation problem is solved to global optimality for each queried x, and that all ordering breakpoints and stationary candidates are enumerated within numerical tolerance.

5. Simulation

In this section, the simulation results are presented to validate the structural findings and evaluate the impact of PASS on FL. Unless otherwise stated, the system parameters used across simulations are listed in Table 1 [10,22]. We consider $N \in \{10, 20, 30\}$ clients with computation times drawn i.i.d. from $t_i^{\mathrm{comp}} \sim \mathrm{Unif}[0.05, 0.15]$ s. To model non-IID data, we synthesize per-client label distributions using a Dirichlet model with concentration $\alpha_{\mathrm{data}} = 0.3$; the heterogeneity score for each client is defined as the ℓ_2 deviation from the global label distribution. We use this score (after positive normalization) as the surrogate heterogeneity weight c_i in $g(q)$ and in the statistics-only baseline. Accordingly, the MNIST/CIFAR-10 experiments below should be interpreted as empirical evaluations of the proxy-based design induced by $J(q,x) = f(q,x)g(q)$, rather than as validations of a convex convergence theorem. We implement synchronous FedAvg for $R = 100$ rounds with per-round participation. In each round, the server selects K participating clients according to a distribution q; K denotes the per-round participation level and is swept as indicated in

each figure. Each selected client performs $E = 2$ local epochs of SGD with batch size 32 and learning rate 0.01 [26]. The server aggregates client models by weighting proportional to local data size. Test accuracy is evaluated every five rounds. For "1× data" and "2× data", each client holds 100 and 200 training samples, respectively. We use a lightweight CNN: for MNIST, two 3×3 conv layers (32 and 64 channels) followed by a 128-unit FC layer; for CIFAR-10, two 3×3 conv layers (64 and 128 channels) with 2×2 max-pooling after each conv, followed by a 256-unit FC layer [27].

Table 1. Simulation parameters.

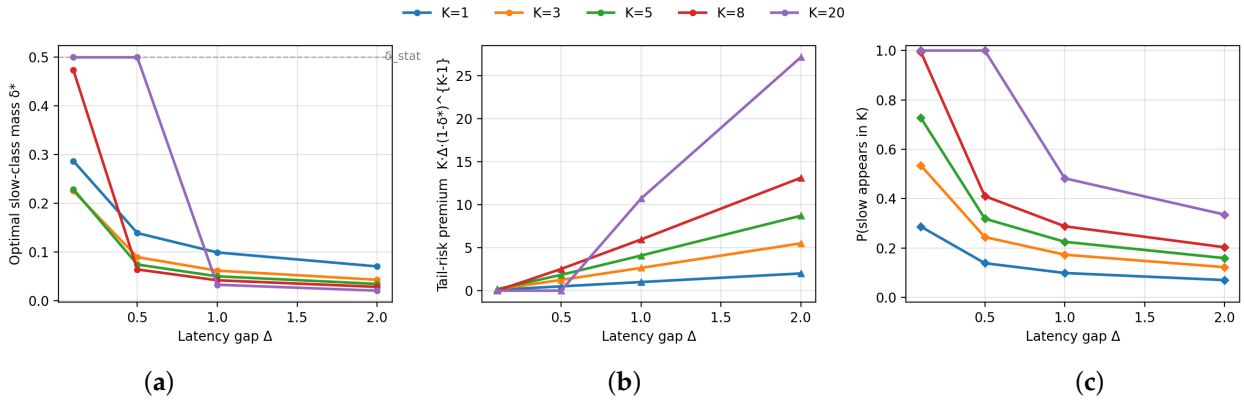
Item	Value
Bandwidth B	10 MHz
Transmit power P	23 dBm
Noise PSD N_0	-174 dBm/Hz
Waveguide length L	10 m
Number of clients N	$\{10, 20, 30\}$
Per-round participants K	varied (see figures)
Compute time t_i^{comp}	Unif$[0.05, 0.15]$ s
Non-IID split	Dirichlet, $\alpha_{\text{data}} = 0.3$
FedAvg rounds R	100
Local training	$E = 2$ epochs, batch size 32, learning rate 0.01
Data per client	100 ("1×") / 200 ("2×") samples
MNIST CNN	3×3 conv (32)–ReLU, 3×3 conv (64)–ReLU, FC 128
CIFAR-10 CNN	3×3 conv (64)–ReLU–pool, 3×3 conv (128)–ReLU–pool, FC 256

Unless otherwise stated, "Conventional" uses fixed placement $x = 0$ to model conventional antenna and uniform participation q_{unif} with $q_i = 1/N$. Placement baselines: (i) fixed placement ($x = 0$); (ii) centroid placement $x_{\text{cen}} = \frac{1}{N}\sum_{i=1}^{N} u_i$; (iii) min–max placement $x_{\min\max} = \arg\min_x \max_i d_i(x)$; (iv) random placement $x \sim \text{Unif}[0, L]$ (averaged over seeds); (v) grid search on a uniform grid over $[0, L]$ to obtain x_{grid} by brute-force evaluation of the relevant objective—for placement-only comparisons, we evaluate $J(q_{\text{unif}}, x)$ (equivalently, $f(q_{\text{unif}}, x)$ since $g(q_{\text{unif}})$ is constant), and, for joint comparisons, we evaluate the envelope $J^\star(x) = \min_{q \in \Delta} J(q, x)$. Participation baselines (for fixed x): (i) fastest-K—pick the K clients with smallest $t_i(x)$ and form $\tilde{q}$ with $\tilde{q}_i = 1/K$ on this set and 0 otherwise; when we report $g(q)$, we use an ε-smoothed distribution $q = (1 - \varepsilon)\tilde{q} + \varepsilon q_{\text{unif}}$; (ii) inverse-latency sampling $q_i = \frac{(t_i(x)+\epsilon_0)^{-1}}{\sum_j (t_j(x)+\epsilon_0)^{-1}}$ with a small $\epsilon_0 > 0$; (iii) statistics-only sampling $q_i = \frac{\sqrt{c_i}}{\sum_j \sqrt{c_j}}$, where c_i is the heterogeneity weight (set to the Dirichlet-based score defined above). In addition, we define PASS-PlacementOnly as optimizing x with fixed $q = q_{\text{unif}}$, and PASS-Joint as jointly optimizing (q, x).

Unless otherwise stated, end-to-end FedAvg experiments use uniform participation $q = q_{\text{unif}}$, so "PASS-optimized placement" refers to PASS-PlacementOnly (optimizing x only).

Figure 1 illustrates how the optimal slow-class mass $\delta^\star$ responds to the latency gap Δ. Figure 1a shows that $\delta^\star$ decreases monotonically with Δ across all K: as the slow–fast gap widens, the wall-clock penalty for sampling slow clients increases, and the optimizer suppresses $\delta^\star$. The decrease is sharper for larger K because the probability of drawing at least one slow client grows with K, making straggling nearly certain unless $\delta^\star$ is reduced. Figure 1b plots the tail-latency premium $K\Delta(1 - \delta^\star)^{K-1}$—the marginal latency cost per unit increase in slow-class mass. As Δ grows and $\delta^\star$ shrinks, $(1 - \delta^\star)^{K-1} \approx 1$, yielding an effective amplification close to $K\Delta$. This is the regime where order-statistics straggling most strongly distorts the allocation: small changes in δ have large wall-clock consequences. Figure 1c shows the probability $1 - (1 - \delta^\star)^K$ that at least one slow client is sampled.

The optimizer implicitly targets a non-saturating straggler probability by shrinking $\delta^\star$ as Δ increases. The scaling $\delta = \rho/K$ yields $(1-\delta)^K \to e^{-\rho}$, explaining why the $\Theta(1/K)$ participation scale emerges naturally.



(**a**) (**b**) (**c**)

Figure 1. Tail-latency premium induced by synchronous straggling. (**a**) Optimal slow-class mass $\delta^\star$ versus latency gap Δ. (**b**) Tail-latency premium $K\Delta(1-\delta^\star)^{K-1}$ at the optimizer. (**c**) Probability of sampling at least one slow client, $1-(1-\delta^\star)^K$. Larger K amplifies the straggler penalty, driving $\delta^\star$ toward the $\Theta(1/K)$ scale.

Figure 2 compares the complementary CDF $\mathbb{P}(T_{\text{round}} > t)$ for conventional fixed placement (dashed) and PASS-optimized placement (solid). Across exceedance probabilities down to $\sim 10^{-6}$, PASS exhibits a systematic left shift, reducing both mean round time and the high quantiles that dominate wall-clock performance under synchronization. The persistent separation deep into the tail indicates that PASS improves the tail structure, reducing extreme straggler rounds rather than merely shifting the mean. As N increases, the conventional tail degrades more noticeably, while PASS remains stable, consistent with PASS mitigating worst-link outcomes that govern the maximum-order statistic.

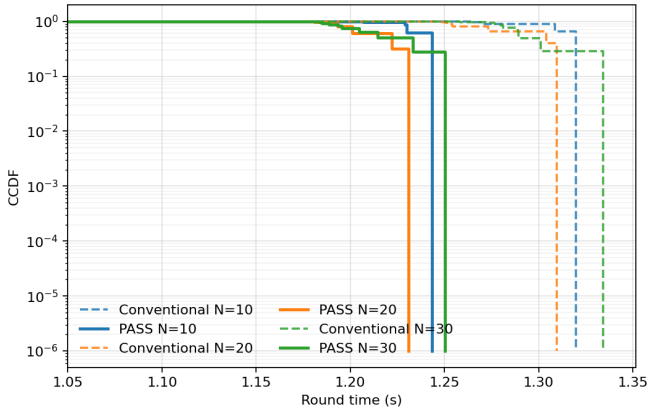


Figure 2. Complementary CDF of synchronous round time T_{round} under conventional fixed placement and PASS-optimized placement, for $N \in \{10, 20, 30\}$ clients. PASS shifts the tail left, reducing extreme straggler rounds that dominate wall-clock training.

Figure 3 validates the phase transition predicted by Theorem 1. Panel (a) shows that, for smaller C_s, $\delta^\star$ decreases rapidly with K and approaches the $1/K$ reference, while, for larger C_s, $\delta^\star$ remains $O(1)$—the slow class's statistical value justifies frequent inclusion despite its latency cost. Panel (b) plots $K\delta^\star$: bounded (flat) behavior confirms $\delta^\star = \Theta(1/K)$, while linear growth indicates $\delta^\star = \Theta(1)$. Panel (c) shows $\delta^\star/\delta^\star_{\text{stat}}$, the ratio to the statistics-only optimum. For small C_s, this ratio decreases with K, indicating that

tail latency increasingly dominates; for large C_s, it remains near unity, indicating statistical considerations prevail.

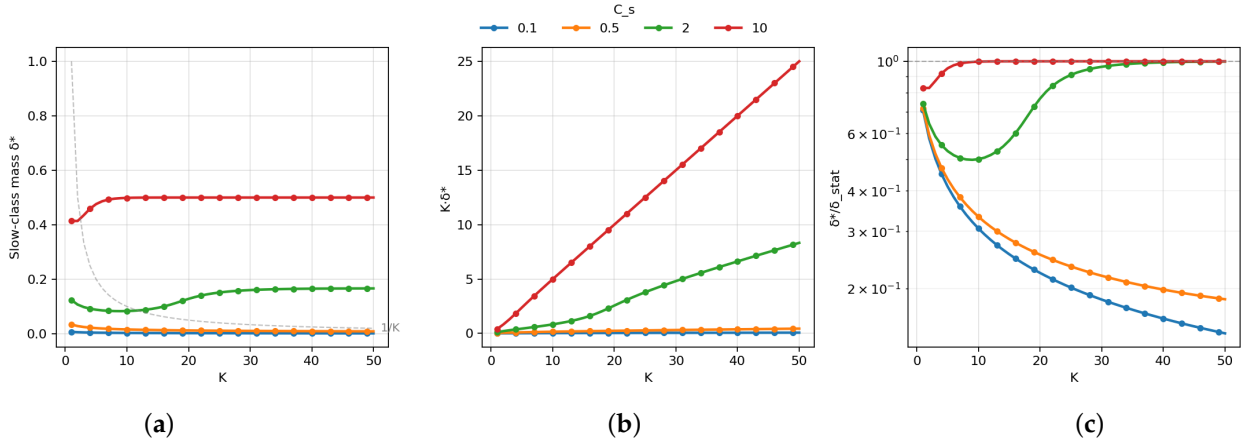


Figure 3. Phase transition of slow-class participation. (**a**) Optimal $\delta^\star$ versus K for different C_s values. (**b**) Normalized product $K\delta^\star$: bounded behavior indicates $\delta^\star = \Theta(1/K)$; linear growth indicates $\delta^\star = \Theta(1)$. (**c**) Ratio $\delta^\star/\delta^\star_{\text{stat}}$, quantifying deviation from the statistics-only optimum.

Figure 4 illustrates the piecewise structure of the placement problem. The upper panel shows per-client latency profiles $t_i(x)$; curve crossings change the latency ordering and induce breakpoints. The lower panel shows the envelope objective $J^\star(x) = \min_{q\in\Delta} J(q, x)$: within any interval where the ordering is fixed, $J^\star(x)$ varies smoothly, but its slope changes at breakpoints where $t_i(x) = t_j(x)$.

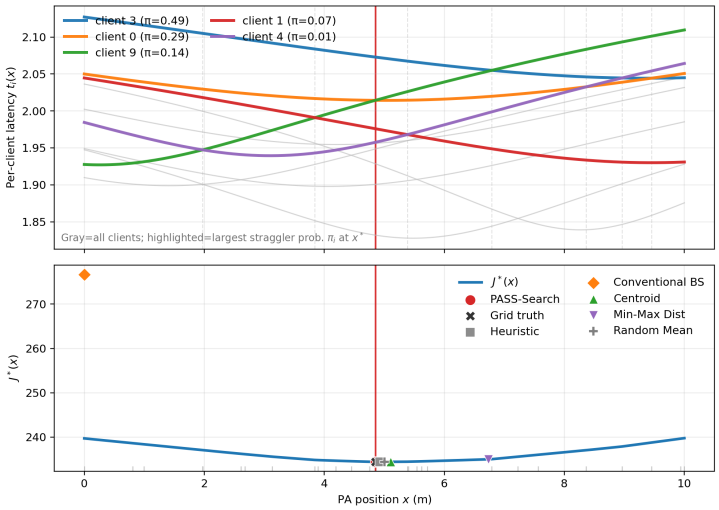


Figure 4. Piecewise-smooth envelope objective versus PA position. (**Top panel**) per-client latency profiles $t_i(x)$; crossings induce ordering changes. (**Bottom panel**) envelope objective $J^\star(x)$; breakpoints occur where $t_i(x) = t_j(x)$. At the optimum $x^\star$ (red line), straggling concentrates on different clients.

At the optimum $x^\star$ (red vertical line), straggling is concentrated on a few clients: $\pi_3 \approx 0.48$, $\pi_0 \approx 0.29$, $\pi_9 \approx 0.14$, $\pi_4 \approx 0.07$, and $\pi_1 \approx 0.01$. This concentration yields an interpretable stationarity condition: optimality requires $\phi(x) = \sum_i \pi_i(x)\, t_i'(x) \approx 0$, i.e., the dominant straggler curves contribute opposing local trends that cancel near $x^\star$.

PASS-Search (Algorithm 1's breakpoint-aware outer search over x; one evaluation computes $J^\star(x)$ at a queried x, including the inner solve for $\hat{q}(x)$) matches the grid-search placement baseline x_{grid} (black $\times$) while requiring far fewer evaluations. In contrast, centroid and min–max placements (geometry-only baselines), random placement $x \sim \text{Unif}[0, L]$, and the conventional baseline at $x = 0$ incur higher $J^\star(x)$, demonstrating that performance

gains require jointly accounting for order-statistics straggling and the breakpoint-induced objective geometry.

Figure 5 visualizes the latency–statistical-efficiency trade-off in the (f, g) plane. Each design point corresponds to a specific method and sample size K: Conventional (fixed placement with uniform sampling), PASS-PlacementOnly (optimized PA placement with uniform sampling), and PASS-Joint (joint optimization of PA placement and participation). The dashed iso-J contours show level sets of the proxy objective $J = f \cdot g$; points closer to the origin lie on lower contours and indicate better wall-clock efficiency under this proxy. Conventional designs cluster in the upper-right region ($f \approx 1.7$, $g \approx 805$), reflecting high straggler latency under fixed placement. PASS-PlacementOnly shifts points left ($f \approx 1.25$) by optimizing PA position, while PASS-Joint further reduces g (to ≈ 780) by jointly tuning participation. Across all K, PASS-Joint consistently occupies the lowest iso-J contours, demonstrating that joint optimization yields Pareto-dominant operating points in the latency–statistical-efficiency space.

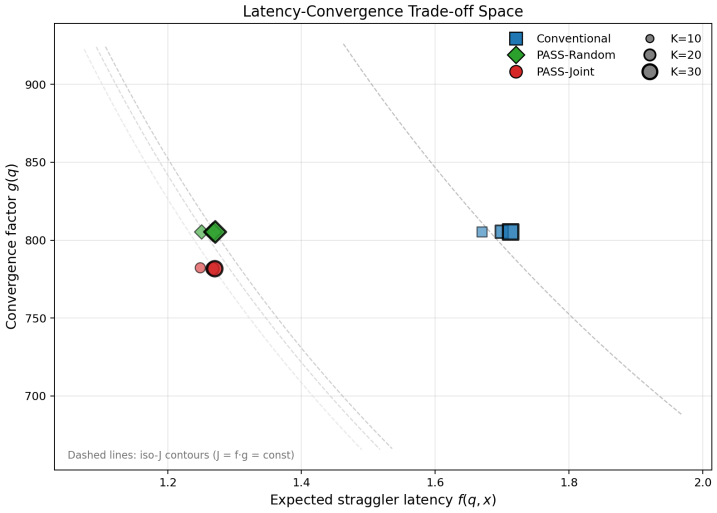


Figure 5. Latency–statistical-efficiency trade-off space. Each marker is a design point in the (f, g) plane, where $f(q, x) = \mathbb{E}[T_{\text{round}}(q, x)]$ (Equation (7)) and $g(q) = \omega \sum_i c_i / q_i + \nu$ is the statistical-efficiency proxy. Marker shape indicates the method (Conventional, PASS-PlacementOnly, PASS-Joint) and marker size indicates the per-round participation level K (varied; see legend).

Figure 6 decomposes the proxy objective into its constituent factors as K varies. The left panel shows that the expected straggler latency $f(q, x)$ increases with K under Conventional placement: as more clients are sampled, the probability of drawing a slow client (and thus incurring the worst-case latency) grows, consistent with the order-statistics analysis in Section 3. In contrast, PASS-Joint maintains a nearly flat f across K by repositioning the PA to shorten the worst links. The center panel shows that scheduling baselines can substantially change the statistical-efficiency proxy $g(q)$: latency-only policies (e.g., fastest-K or inverse-latency sampling) reduce f but typically increase g under non-IID data due to biased participation, whereas statistics-only sampling reduces g but may increase f. The right panel plots the product $J = f \cdot g$: Conventional exhibits a 20–30% increase in J as K grows from 1 to 50, while PASS-Joint remains stable, confirming that PASS mitigates the K-amplified tail penalty identified theoretically.

Figure 7 evaluates end-to-end FL training under a wall-clock budget, comparing Conventional and PASS on MNIST and CIFAR-10. Each curve tracks test accuracy as training progresses in real time, with solid lines indicating PASS-optimized placement and dashed lines indicating conventional fixed placement. On MNIST (Figure 7a), PASS (2× data) reaches approximately 95% accuracy within 20 s and maintains this level, whereas

Conventional (2× data) plateaus around 77%. The gap is even more pronounced on CIFAR-10 (Figure 7b), where PASS achieves around 61% accuracy compared to around 49% for Conventional under the same data and time budget. These results confirm that the latency reduction provided by PASS translates directly into faster wall-clock learning progress and higher final accuracy under synchronous FL with deadline constraints: by reducing straggler rounds, PASS completes more communication rounds within the same wall-clock window, accumulating greater empirical progress.

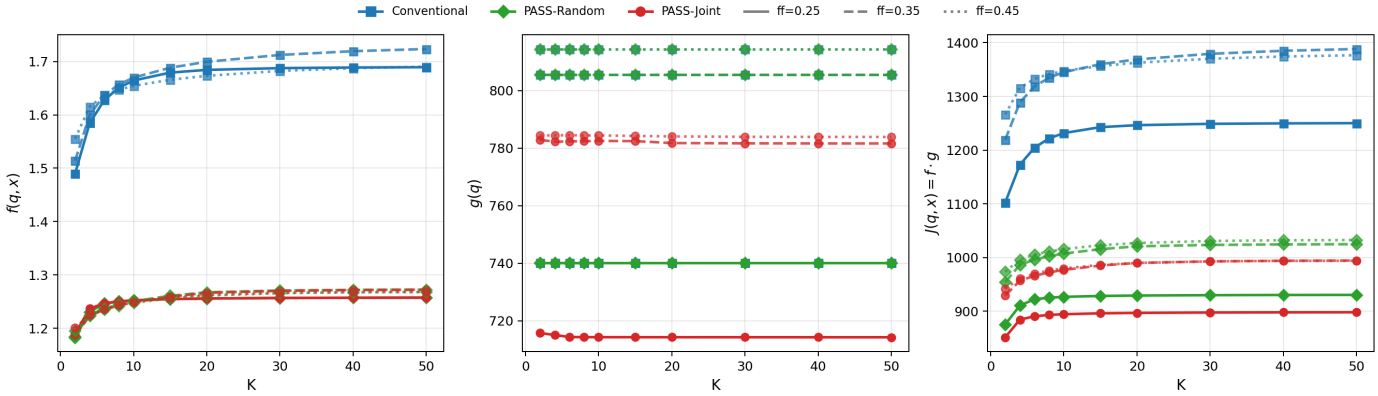
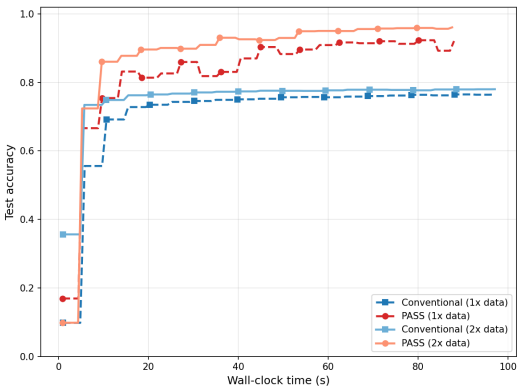
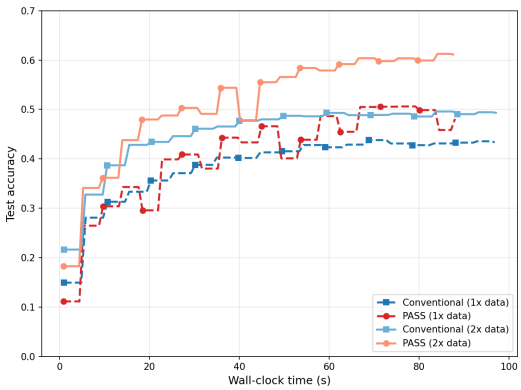


Figure 6. Decomposition of the proxy objective $J = f \cdot g$ versus sample size K, under different fast-class fractions $f_f \in \{0.25, 0.35, 0.45\}$ (fraction of clients in the fast-latency class in the two-class model). (**Left**): expected straggler latency $f(q, x)$. (**Center**): statistical-efficiency proxy $g(q)$. (**Right**): product $J(q, x) = f \cdot g$. Conventional placement exhibits increasing f with K due to order-statistics amplification; PASS-Joint maintains nearly flat f by optimizing PA position. The proxy term g varies relatively mildly across methods. Consequently, the product J grows significantly for Conventional but remains low and stable for PASS-Joint.



(**a**) MNIST.



(**b**) CIFAR-10.

Figure 7. Test accuracy versus wall-clock time under deadline-constrained synchronous FL. Solid lines denote PASS-optimized placement; dashed lines denote conventional fixed placement. "1× data" and "2× data" indicate the relative local dataset size per client. On both datasets, PASS achieves higher accuracy at any given wall-clock budget, with the gap widening under increased data (where more rounds are needed to reach a target accuracy). PASS (2× data) reaches approximately 95% on MNIST (**a**) and 61% on CIFAR-10 (**b**), compared to 77% and 49% for Conventional (2× data), respectively.

6. Conclusions

This paper developed a joint theory of client participation and pinching-antenna placement for PASS-enabled FL under non-IID data. The KKT recursion reveals how latency gaps induce K-amplified tail penalties in the optimal sampling distribution, distorting the classical square-root rule. Under latency-class structure, we established within-class

square-root sampling and derived a two-class phase-transition threshold under which slow-class probability mass collapses to the $O(1/K)$ scale. For antenna placement, we characterized the piecewise envelope structure and provided an exact breakpoint-and-root candidate-enumeration algorithm. Simulations confirm that the joint optimization yields Pareto-dominant operating points in the latency–statistical-efficiency trade-off and translates to substantial accuracy gains. Future directions include multi-PA architectures with coordinated placement across segmented waveguides, and asynchronous FL where clients contribute under bounded staleness rather than strict synchronization.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L. and Z.D.; Software, Y.L.; Validation, Y.L.; Formal analysis, Y.L.; Writing—original draft, Y.L.; Writing—review and editing, Y.L. and Z.D.; Supervision, Z.D.; Funding acquisition, Z.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Lemma 1

For fixed x, the per-round latencies are positive constants with $0 < t_1 \le t_i \le t_N$. Hence, the straggler expectation satisfies $t_1 \le f(q) \le t_N$ for all $q \in \Delta$. Moreover, $g(q) = \omega \sum_i c_i/q_i + \nu \ge \omega c_s/q_s$ for any index s. Thus, if $q_s \to 0^+$, then $g(q) \to +\infty$ and $J(q) \ge t_1 g(q) \to +\infty$.

Appendix B. Proof of Lemma 2

First, consider the function $\phi(u) = u^K$ on $u \ge 0$. For integer $K \ge 1$, $\phi''(u) = K(K-1)u^{K-2} \ge 0$, so ϕ is convex on $[0, \infty)$. $Q_i(q) = \sum_{j=1}^{i} q_j$ is affine in q; hence, $Q_i(q)^K = \phi(Q_i(q))$ is convex in q. Since $\Delta_i \ge 0$, the term $-\Delta_i Q_i(q)^K$ is concave. Summing over $i = 1, \ldots, N-1$ and adding the constant t_N establishes the concavity of f on $\bar{\Delta}$.

Second, fix $q \in \Delta$ and a feasible direction v with $\sum_{i=1}^{N} v_i = 0$. Then, $\mathrm{D}Q_i(q)[v] = \sum_{j=1}^{i} v_j$. Applying the chain rule, we obtain $\mathrm{D}f(q)[v] = -\sum_{i=1}^{N-1} \Delta_i \cdot K Q_i^{K-1} \cdot \sum_{j=1}^{i} v_j$. Reordering this double sum yields $\mathrm{D}f(q)[v] = -\sum_{s=1}^{N} \left(K \sum_{i=s}^{N-1} \Delta_i Q_i^{K-1} \right) v_s = -\sum_{s=1}^{N} D_s(q)\, v_s$, which confirms (17).

Finally, for $K \ge 2$, differentiating (17) once more yields the Hessian representation in (18). Since each $a_i a_i^\top \succeq 0$ and each coefficient $\Delta_i Q_i^{K-2} \ge 0$, we have $\nabla^2 f(q) \preceq 0$, consistent with the concavity property derived in (i).

Appendix C. Proof of Proposition 2

In the sorted order, all fast clients precede slow clients. There is exactly one nonzero gap at the class boundary: $\Delta_{n_f} = \Delta$, where $n_f \triangleq |\mathcal{F}|$. Thus, for any slow index $s > n_f$, $D_s = 0$; for any fast index $s \le n_f$, $D_s = K\Delta Q_{n_f}^{K-1} = K\Delta(1-\delta)^{K-1}$.

By the within-class square-root form, for any fast client i, $c_i/(q_i^\star)^2 = C_f^2/(1-\delta^\star)^2$, and, for any slow client j, $c_j/(q_j^\star)^2 = C_s^2/(\delta^\star)^2$. Applying (19) to any slow client gives $\lambda = f\omega C_s^2/(\delta^\star)^2$. Applying (19) to any fast client and substituting λ yields (29).

Let $\bar{\Delta} = \{q \in \mathbb{R}^N : q_i \ge 0,\ \sum_i q_i = 1\}$. Since $J(q) \to +\infty$ near the boundary of $\bar{\Delta}$, any minimizing sequence is eventually contained in a compact subset $\{q \in \bar{\Delta} : q_i \ge \varepsilon\ \forall i\}$ for some $\varepsilon > 0$. By compactness and continuity of J on this subset, a global minimizer exists and must satisfy $q_i > 0$ for all i, i.e., $q^{\mathrm{opt}} \in \Delta$.

Appendix D. Proof of Theorem 1

The key observation is that the tail term $(1-\delta)^K$ becomes negligible as soon as $K\delta \to \infty$. Assume, for contradiction, that there exists a subsequence $\{K_n\}$ such that the corresponding minimizers satisfy $K_n\delta^\star_{K_n} \to \infty$. Then, $(1-\delta^\star_{K_n})^{K_n} \le \exp\{-K_n\delta^\star_{K_n}\} \to 0$ and, hence,

$$f(\delta^\star_{K_n}) = t_s - \Delta(1-\delta^\star_{K_n})^{K_n} \to t_s. \tag{A1}$$

On the other hand, for all $\delta \in (0,1)$, we have the uniform lower bound

$$g(\delta) = \omega\Big(\frac{C_f^2}{1-\delta} + \frac{C_s^2}{\delta}\Big) + \nu \ge \omega C_f^2 + \nu. \tag{A2}$$

Therefore,

$$\liminf_{n\to\infty} J_{K_n}(\delta^\star_{K_n}) = \liminf_{n\to\infty} f(\delta^\star_{K_n}) g(\delta^\star_{K_n}) \ge t_s(\omega C_f^2 + \nu). \tag{A3}$$

Now, evaluate the objective at the tail test point $\delta_\rho = \rho/K$:

$$J_K(\delta_\rho) = \big[t_s - \Delta P_{\rho,K}\big]\left[\omega\Big(\frac{C_f^2}{1-\rho/K} + \frac{C_s^2 K}{\rho}\Big) + \nu\right]. \tag{A4}$$

Using (31), we bound the slow-class term as $\omega(C_s^2 K/\rho) \le (1-\xi)(\omega C_f^2 + \nu)\,\Delta P_{\rho,K}/(t_s - \Delta P_{\rho,K})$. Substituting this into (A4) yields

$$J_K(\delta_\rho) \le \big[t_s - \Delta P_{\rho,K}\big]\Big[\omega\tfrac{C_f^2}{1-\rho/K} + \nu\Big] + (1-\xi)(\omega C_f^2 + \nu)\,\Delta P_{\rho,K}. \tag{A5}$$

Since ρ is fixed and $K \to \infty$, we have $\frac{1}{1-\rho/K} = 1 + O(1/K)$ and $P_{\rho,K} \to e^{-\rho}$. Consequently, there exists K_0 such that, for all $K \ge K_0$,

$$J_K(\delta_\rho) \le t_s(\omega C_f^2 + \nu) - \tfrac{\xi}{2}(\omega C_f^2 + \nu)\,\Delta P_{\rho,K}. \tag{A6}$$

Combining (A6) with (A3) contradicts the optimality of $\delta^\star_{K_n}$ for all sufficiently large n. Hence, $K\delta^\star_K$ cannot diverge; equivalently, $\sup_K K\delta^\star_K < \infty$, i.e., $\delta^\star_K = O(1/K)$.

References

1. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017.
2. Yang, Z.; Chen, M.; Wong, K.K.; Poor, H.V.; Cui, S. Federated Learning for 6G: Applications, Challenges, and Opportunities. *Engineering* **2022**, *8*, 33–41. [CrossRef]
3. Qin, Z.; Li, G.Y.; Ye, H. Federated Learning and Wireless Communications. *IEEE Wirel. Commun.* **2021**, *28*, 134–140. [CrossRef]
4. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated Learning with Non-IID Data. *arXiv* **2018**, arXiv:1806.00582. [CrossRef]
5. Samy, M.; Al-Hraishawi, H.; Alsenwi, M.; Adam, A.B.M.; Chatzinotas, S.; Ottersten, B. Pinching Antenna Systems Versus Reconfigurable Intelligent Surfaces in mmWave. *IEEE Wirel. Commun. Lett.* **2025**, *14*, 4022–4026. [CrossRef]
6. Ding, Z.; Schober, R.; Poor, H.V. Flexible-Antenna Systems: A Pinching-Antenna Perspective. *IEEE Trans. Commun.* **2025**, *73*, 9236–9253. [CrossRef]
7. Zhu, G.; Mu, X.; Guo, L.; Xu, S.; Liu, Y.; Al-Dhahir, N. Pinching-Antenna Systems (PASS)-Enabled Secure Wireless Communications. *IEEE Trans. Commun.* **2026**, *74*, 490–505. [CrossRef]
8. He, J.; Mu, X.; Ngo, H.Q.; Matthaiou, M. Pinching-Antenna System-Assisted Localization: A Stochastic Geometry Perspective. *IEEE Wirel. Commun. Lett.* **2026**, *15*, 1737–1741. [CrossRef]
9. Ai, Y.; Mu, X.; Si, P.; Liu, Y. Delay Minimization in Pinching-Antenna-Enabled NOMA-MEC Networks. *IEEE Commun. Lett.* **2026**, *30*, 962–966. [CrossRef]

10. Ding, Z.; Poor, H.V. Analytical Optimization for Antenna Placement in Pinching-Antenna Systems. *arXiv* **2025**, arXiv:2507.13307. [CrossRef]
11. Zhao, J.; Song, H.; Mu, X.; Cai, K.; Zhu, Y.; Liu, Y. Pinching-antenna systems-enabled multi-user communications: transmission structures and beamforming optimization. *IEEE Trans. Commun.* **2025**, *advance online publication*. [CrossRef]
12. Li, H.; Zhong, R.; Pan, Z.; Dong, C.; Lei, J.; Liu, Y. Pinching Antenna Systems for Integrated Sensing and Communications. *arXiv* **2025**, arXiv:2508.19540. [CrossRef]
13. Lai, F.; Zhu, X.; Madhyastha, H.V.; Chowdhury, M. Oort: Efficient Federated Learning via Guided Participant Selection. In Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21), Virtual, 14–16 July 2021; pp. 19–35.
14. Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; Cheng, Y. TiFL: A Tier-based Federated Learning System. In Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '20), Stockholm, Sweden, 23–26 June 2020; pp. 125–136. [CrossRef]
15. Nguyen, J.; Malik, K.; Zhan, H.; Yousefpour, A.; Rabbat, M.; Malek, M.; Huba, D. Federated Learning with Buffered Asynchronous Aggregation. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS), Virtual, 28–30 March 2022; Volume 151, pp. 3581–3607.
16. Liu, Y.J.; Feng, G.; Du, H.; Qin, Z.; Sun, Y.; Kang, J.; Li, X.; Niyato, D. Adaptive Clustering-Based Straggler-Aware Federated Learning in Wireless Edge Networks. *IEEE Trans. Commun.* **2024**, *72*, 7757–7771. [CrossRef]
17. Reisizadeh, A.; Tziotis, I.; Hassani, H.; Mokhtari, A.; Pedarsani, R. Straggler-Resilient Federated Learning: Leveraging the Interplay Between Statistical Accuracy and System Heterogeneity. *IEEE J. Sel. Areas Inf. Theory* **2022**, *3*, 197–205. [CrossRef]
18. Pan, C.; Ren, H.; Wang, K.; Kolb, J.F.; Elkashlan, M.; Chen, M.; Di Renzo, M.; Hao, Y.; Wang, J.; Swindlehurst, A.L.; et al. Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions. *IEEE Commun. Mag.* **2021**, *59*, 14–20. [CrossRef]
19. Xiao, Z.; Pi, X.; Zhu, L.; Xia, X.G.; Zhang, R. Multiuser Communications With Movable-Antenna Base Station: Joint Antenna Positioning, Receive Combining, and Power Control. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 19744–19759. [CrossRef]
20. Tyrovolas, D.; Tegos, S.A.; Diamantoulakis, P.D.; Ioannidis, S.; Liaskos, C.K.; Karagiannidis, G.K. Performance Analysis of Pinching-Antenna Systems. *IEEE Trans. Cogn. Commun. Netw.* **2026**, *12*, 590–601. [CrossRef]
21. Lyu, Z.; Li, H.; Gao, Y.; Xiao, M.; Poor, H.V. Pinching-Antenna Systems (PASS) Aided Over-the-Air Computation. *IEEE Commun. Lett.* **2025**, *29*, 2531–2535. [CrossRef]
22. Lin, Y.; Chen, Z.; Ding, Z. Pinching-antenna-enabled Federated Learning: Tail Latency, Participation, and Convergence Analysis. *arXiv* **2025**, arXiv:2510.23315. [CrossRef]
23. Wu, B.; Fang, F.; Zeng, M.; Wang, X. Straggler-Resilient Federated Learning over A Hybrid Conventional and Pinching Antenna Network. *arXiv* **2025**, arXiv:2508.15821. [CrossRef]
24. Chen, M.; Yang, Z.; Saad, W.; Yin, C.; Poor, H.V.; Cui, S. A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 269–283. [CrossRef]
25. Yang, Z.; Chen, M.; Saad, W.; Hong, C.S.; Shikh-Bahaei, M. Energy Efficient Federated Learning Over Wireless Communication Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 1935–1949. [CrossRef]
26. Lin, Y.; Wang, K.; Ding, Z. Rethinking Clustered Federated Learning in NOMA Enhanced Wireless Networks. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 16875–16890. [CrossRef]
27. Wang, K.; Ma, Y.; Mashhadi, M.B.; Foh, C.H.; Tafazolli, R.; Ding, Z. Convergence Acceleration in Wireless Federated Learning: A Stackelberg Game Approach. *IEEE Trans. Veh. Technol.* **2025**, *74*, 714–729. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.