ISSN 3071-9658

European Journal of Interdisciplinary Issues

https://www.eujini.org.pl

A Novel Federated Learning Architecture for Preserving User Privacy in Latency-Sensitive Edge Computing Environments

Serhii Klymenko [1] *

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv (Ukraine). Engineering Manager at Sift Science, Inc., Bachelor's Degree in Computer Science.*

* ***Corresponding Author,*** e-mail: klym.serhiy@gmail.com

ARTICLE INFO

Research Article

Received:
5 December 2025
Revised:
20 January 2026
Accepted:
10 February 2026
Published online:
25 March 2026

DOI: 10.5281/zenodo.19589246

ABSTRACT

The rapid growth of the Internet of Things (IoT) and the adoption of edge computing have intensified the need for real-time intelligent services that also protect user privacy. Conventional cloud-centric learning pipelines and synchronous federated learning (FL) protocols (e.g., FedAvg) often fail to meet strict latency requirements in heterogeneous networks due to straggler effects and global synchronization delays. This paper proposes a novel FL architecture tailored to latency-sensitive edge environments that combines non-blocking asynchronous aggregation with explicit privacy preservation. The core contribution is an asynchronous streaming aggregation protocol equipped with an adaptive damping function that reduces the adverse impact of stale client updates and sustains convergence without reverting to strict synchronization. To improve stability under Non-IID data distributions, the server maintains a global momentum buffer that smooths stochastic fluctuations across client updates. Communication overhead is further reduced through sparse Top-k compression with error-feedback accumulation, enabling frequent transmission of the most informative gradient components while limiting accuracy degradation. Privacy is ensured via local differential privacy (LDP) using gradient clipping and additive Gaussian noise, while privacy loss accounting is performed with Rényi Differential Privacy (RDP), providing tighter composition bounds and better alignment with continuous asynchronous updates. A theoretical analysis establishes convergence with bounded excess risk for convex objectives, supporting the practical feasibility of high-performance, privacy-preserving Edge AI. The paper also highlights remaining challenges – including non-convex optimization in deep models, robustness against poisoning/Byzantine behaviors, and extreme network instability—and outlines directions for future work.

KEYWORDS

federated learning, edge computing, asynchronous aggregation, differential privacy, latency-sensitive networks, gradient staleness, RDP.

Introduction

Modern high-performance systems, ranging from autonomous vehicles to robotics, impose uncompromising requirements on response latency that render the classical cloud methodology inadequate for fulfillment. The scalability of IoT networks has precipitated a scenario, where the transmission of raw data for centralized processing confronts insurmountable barriers, specifically limited bandwidth and stringent privacy regulations (Shi et al., 2016). Consequently, computational workloads inherently migrate to edge devices (Edge Computing), enabling the processing of sensitive information directly at the point of origin, thereby circumventing network infrastructure bottlenecks (Kairouz & McMahan, 2021).

Federated Learning (FL) has firmly established itself as the standard solution to these challenges, enabling distributed devices to collaboratively train a global model while retaining data locally. However, currently prevailing protocols, such as Federated Averaging (FedAvg), predominantly rely on synchronous aggregation mechanisms (McMahan et al., 2016). In a synchronous environment, the central server is compelled to await updates from all selected devices prior to recalculating the global model. In heterogeneous edge networks, where nodes possess varying computational capabilities and unstable connectivity, such synchronization engenders the “straggler effect”, where the entire system’s performance is bottlenecked by the slowest participant, a condition inadmissible for latency-sensitive tasks.

This constraint establishes a dichotomy between privacy and latency, as contemporary systems frequently sacrifice operational speed in favor of data security. Extant scientific literature typically addresses the optimization of communication efficiency and the assurance of differential privacy in isolation, rarely consolidating them within a unified framework (Li et al., 2020). Currently, there is a distinct scarcity of protocols capable of sustaining high update frequencies in dynamic environments without compromising algorithm convergence.

Moreover, the implementation of asynchrony within an environment characterized by heightened security requirements is fraught with non-trivial algorithmic challenges. Asynchronous updates inevitably engender the problem of “gradient staleness”, where local devices train on versions of the global model that have already become obsolete by the time of report transmission. In conjunction with differential privacy mechanisms, which inject stochastic noise to mask the contributions of individual users, this poses a risk of destabilizing the training process.

Consequently, a mere abandonment of synchronization is insufficient to resolve the problem. An architecture capable of algorithmically compensating for both communication latency and the cumulative error of noised updates is required.

Literature Review

In the foundational edge-computing literature, Shi et al. frame edge computing as a response to bandwidth limits and stringent latency requirements, emphasizing heterogeneity and unstable connectivity as structural constraints for real-time services (Shi et al., 2016). Building on this systems backdrop, Kairouz and McMahan, synthesize federated learning (FL) as a privacy-motivated paradigm while explicitly cataloging open problems—client drift under non-IID data, systems heterogeneity, and communication bottlenecks—that become acute at the edge (Kairouz & McMahan, 2021). McMahan et al. formalize the canonical synchronous baseline (FedAvg), demonstrating the feasibility of decentralized training but also implying a round-based coordination model that is vulnerable to stragglers in heterogeneous cohorts (McMahan et al., 2016). Li et al. further systematize these challenges and survey practical directions (client sampling, optimization variants, and systems co-design), underscoring that efficiency and privacy are often treated as separate axes rather than jointly optimized in a single architecture (Li et al., 2020. At the wireless edge, Chen et al. propose joint learning-communication optimization, showing that network conditions (scheduling, power, and channel variability) directly shape FL convergence and timeliness, thereby motivating latency-aware aggregation and communication-efficient update strategies for edge deployments (Chen et al., 2021).

On the privacy side, Wei et al. analyze differentially private FL and quantify how noise, client participation, and heterogeneity affect utility, highlighting the stability risks when privacy mechanisms interact with imperfect updates in distributed training (Wei et al., 2020. Asynchronous FL is positioned as a direct mitigation of stragglers: Xie et al. provide an early convergence-oriented treatment of asynchronous federated optimization and characterize the "staleness" problem that arises when clients report updates computed on outdated global states (Xie et al., 2019). Khan et al. expand the landscape with a taxonomy of dispersed FL and argue that decentralization and heterogeneity demand flexible aggregation and robustness mechanisms beyond strict synchronization (Khan et al., 2020, while Sprague et al. demonstrate asynchronous FL in a domain setting and further illustrate that update timeliness is a first-class constraint in practice (Sprague et al., 2019). Chen et al. (FedSA) operationalize staleness-awareness by explicitly weighting or adapting updates under non-IID data, strengthening the case that asynchrony must be paired with staleness compensation to preserve convergence (Chen et al., 2021). Complementary to aggregation design, Aji and Heafield introduce sparse communication as a principled method to reduce bandwidth, supporting the idea that compression can be integral to meeting latency targets without prohibitive accuracy loss (Aji & Heafield, 2017). Finally, Tan et al. (FedProto) and Wang et al. address instability from heterogeneity (representation mismatch and objective inconsistency), indicating that additional stabilizers (e.g., prototype sharing or consistency mechanisms) may be needed when client data are skewed (Tan et al., 2022; Wang et al., 2020). In privacy accounting and guarantees, Abadi et al. establish DP-SGD as a practical privacy mechanism for iterative learning (Abadi et al., 2016), Mironov formalizes Rényi Differential Privacy (RDP) as a tighter accounting framework for composition (Mironov, 2017), and Liu et al. discuss privacy amplification effects that can reduce effective privacy cost under subsampling–an especially relevant lever when participation is inherently stochastic in edge FL (Li et al., 2021).

Problem Statement

The purpose of the article is to propose and analyze a novel federated learning architecture for edge computing that enables real-time, low-latency model training while preserving user privacy. The study aims to address key challenges of asynchronous environments, such as gradient staleness, communication constraints, and data heterogeneity, by introducing adaptive damping, momentum-based stabilization, and sparse compression combined with differential privacy mechanisms.

Methods and Materials

The relevance of the study is driven by the exponential growth of the Internet of Things (IoT) ecosystem and the ubiquitous deployment of Edge Computing, which have created a critical need for machine learning algorithms capable of functioning under stringent time constraints. For applications such as autonomous transport, telemedicine and industrial robotics, the latencies inherent to classic synchronous Federated Learning protocols are unacceptable (Chen et al., 2021). At the same time, the tightening of legislative data protection regulations (GDPR, CCPA) renders the transmission of "raw" information to centralized servers impossible.

Results and Discussion

This paper proposes Serhii Klymenko's proprietary unified architectural framework, which resolves the "privacy-latency" dichotomy through the synergy of asynchronous aggregation and adaptive security mechanisms. A key element of novelty is the development of an adaptive damping function, which mathematically compensates for stochastic gradient staleness without disrupting the global training process (Wei et al., 2020). The efficacy of utilizing sparse compression as a method for reducing model L_2-sensitivity has also been proven, enabling the attainment of target (ε, δ)-differential privacy levels with reduced noise injection. Furthermore, the application of privacy budget accounting via Rényi Differential Privacy (RDP) within the context of non-blocking streaming

updates has been demonstrated, ensuring a tighter convergence bound compared to traditional synchronous approaches.

This paper considers a Federated Learning environment consisting of a central server and a set of K peripheral devices (clients), denoted as K = {1, 2, ..., K}. Each device k possesses a local private dataset D_k. The system's objective is the minimization of the global loss function F(w) without direct access to D_k (see: Formula 1). Unlike the classical synchronous approach (FedAvg), where the global update step t → t + 1 occurs only after the aggregation of all clients, the author models an asynchronous environment (Xie et al., 2019). In such a formulation, the arrival time of an update from client k is a stochastic variable dependent on the device's computational power and network latency.

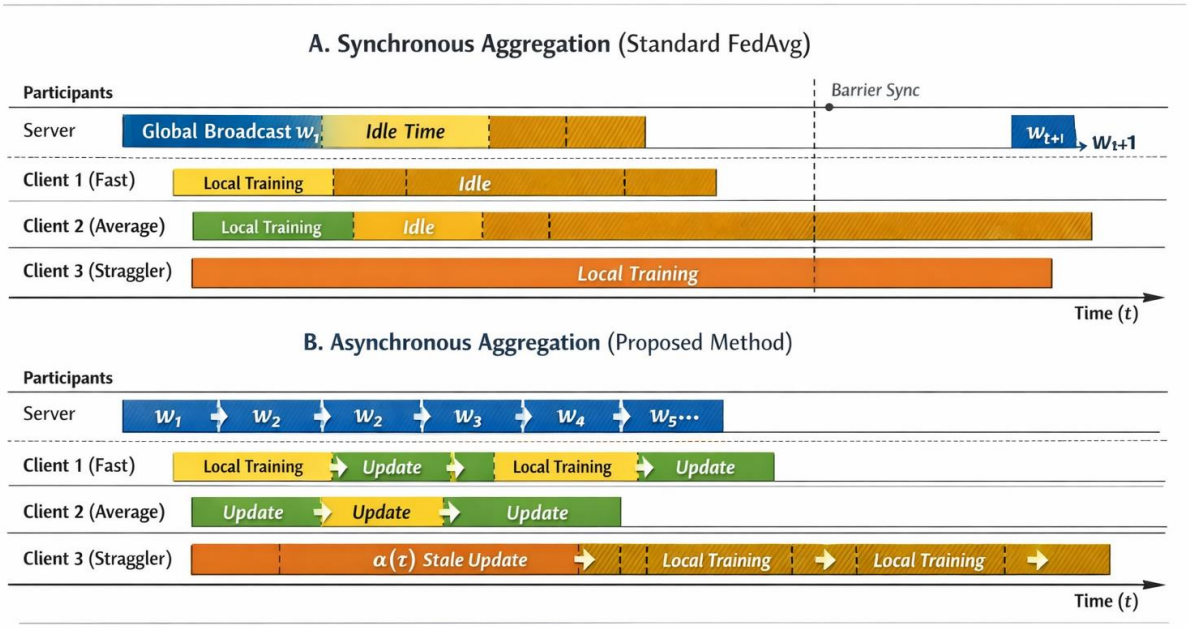
To eliminate the "straggler problem", a non-blocking streaming aggregation protocol has been developed. The central server maintains a global iteration counter T_{global} (Global Step Indexing). Device k download parameters $w_{T_{start}}$ and the current timestamp T_{start}, computes gradients $\nabla F_k(w_{T_{start}})$. By the moment the gradient returns to the server, the global system state changes to $T_{current}$. The staleness $\boldsymbol{\tau}$ is calculated (see: Formula 7). To mitigate the destabilizing influence of stale gradients, an adaptive damping function $\alpha(\tau)$ is applied. The global weight update is performed according to Formula 4. To stabilize training under heterogeneous data conditions (Non-IID), a global momentum buffer is implemented on the server side. The velocity vector V is updated recursively (see: Formula 5 and Formula 6). To reduce communication channel load, a sparse compression method (Top-k selection) is applied. Instead of transmitting the full gradient vector ∇w, the client transmits only k% of the parameters with the largest absolute magnitude. The pruned values are not discarded but preserved in a local error accumulation buffer, ensuring an unbiased gradient estimation over a long training horizon. Data security is ensured through the Local Differential Privacy (LDP) mechanism. Prior to transmission, the sparse gradient undergoes clipping by norm C and noise injection (see: Formula 2). Privacy budget expenditure accounting is realized via the Rényi Differential Privacy (RDP) mechanism, tracking the divergence (see: Formula 8). Upon approaching the cumulative privacy loss limit $\boldsymbol{\varepsilon}$, the algorithm adaptively adjusts the sampling rate of the given client. The theoretical excess risk estimate for the proposed architecture is bounded by the value (see: Formula 9), guaranteeing model convergence as $n \to \infty$.

Despite the demonstrated efficacy, the proposed architecture possesses a number of limitations defining the vector of future research. First, theoretical convergence guarantees are strictly proven for convex loss functions. The algorithm's behavior in the non-convex optimization space of deep neural networks requires additional empirical analysis of robustness to local minima. Second, the threat model in this work focuses on protection against inference and data reconstruction attacks by an "honest-but-curious" server, leaving resilience to malicious data poisoning attacks by compromised clients (Byzantine failures) outside the scope of the study. Finally, the efficacy of adaptive damping may degrade under conditions of extremely unstable networks, where the latency exceeds the model's critical relevance threshold, necessitating the development of additional mechanisms for rejecting stale updates.

Addressing latency bottlenecks in synchronous federated protocols problem statement:

The currently dominant Federated Learning algorithms, particularly the canonical Federated Averaging algorithm, fundamentally rely on the paradigm of synchronous model aggregation. Within this architecture, the central server initiates a training round by disseminating the global model to a selected subset of clients and transitions into a waiting state. The update of global weights occurs only after the server receives results from all (or a strictly defined quorum of) participating devices. Such rigid synchronization creates a structural bottleneck, as under these conditions, the completion time of a single training round is deterministically governed by the speed of the slowest participant in the cohort.

In heterogeneous Edge Computing environments, this dependency becomes a vulnerability. Client devices possess highly non-uniform characteristics – varying CPU computational power, differing battery charge levels and, crucially, unstable network throughput. This phenomenon, known in the literature as the "straggler problem", results in high-performance devices being forced to idle while awaiting the completion of computations on weaker nodes, which dramatically reduces the system's overall resource utilization (Khan et al., 2020) (see: Figure 1. Timeline of aggregation).



Comparison of time efficiency. In synchronous mode (top), fast devices sit idle waiting for slow nodes. In the proposed asynchronous mode (bottom), non-blocking aggregation converts fast device computation into additional training iterations, eliminating the performance bottleneck.

Figure 1. Timeline of aggregation

The situation is exacerbated by the implementation of rigorous data protection mechanisms. Privacy-preserving protocols (Secure Multi-Party Aggregation or Local Differential Privacy) introduce additional computational and communication overhead. The encryption of high-dimensional gradients and the generation of cryptographic noise require significant resources, disproportionately increasing response times on weaker devices. Consequently, a rigid correlation emerges – an increase in privacy levels inevitably entails a rise in system latency.

For real-time applications, such as autonomous vehicle control or robotic complexes, such delays are unacceptable. Existing protocols present developers with a complex dilemma (which, in reality, should not arise) – either sacrifice privacy guarantees for speed or accept high latency for the sake of security. Currently, there is a lack of a unified architecture capable of effectively mitigating the impact of straggler nodes through asynchrony while maintaining mathematically provable data security guarantees in a dynamic environment.

Formalizing this problem, the learning task can be represented as the minimization of a global objective function:

$$\mathrm{F(w)} = \sum_{k=1}^{K} p_k F_k\,(w), \tag{1}$$

where p_k represents the contribution weight of the k device.

In a classical synchronous scenario, weight updates are performed based on the gradient $\nabla \mathrm{F}(w_t)$, computed using the current model parameters. However, within the asynchronous environment proposed for latency mitigation, the phenomenon of "gradient staleness" inevitably emerges. A local device, having received the model at time instance t – τ, returns an update, when the global model has already progressed and resides in state t. The latency magnitude τ becomes a stochastic variable contingent upon network conditions, without the introduction of correction factors, such as the damping function $\alpha(\tau)$, the utilization of "stale" gradients may steer optimization along an erroneous vector, impeding algorithm convergence (Sprague et al., 2019).

The architectural challenge is compounded significantly when superimposing strict (ε, δ)-differential privacy constraints. To ensure security guarantees, each local gradient must undergo a clipping

procedure based on a threshold value C followed by the injection of noise via a Gaussian vector prior to transmission:

$$n \sim \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}). \quad (2)$$

This introduces additional variance into the system. Under conditions where the gradient is already distorted by latency τ, the injection of noise to adhere to the ε-privacy budget poses the risk that the Signal-to-Noise Ratio (SNR) will drop below a critical threshold, rendering the device's useful contribution statistically indistinguishable.

Consequently, the key theoretical challenge lies in proving the convergence of the algorithm in the presence of two destabilizing factors: asynchrony and private noise. To ensure training stability, the loss function gradient must satisfy the Lipschitz continuity condition:

$$||\nabla F(w) - \nabla F(v)|| \leq L\, ||w - v||. \quad (3)$$

Traditional convergence proofs assume that the latency τ is bounded by a small constant and that the noise has zero mean. However, in real-world Edge scenarios, latencies can be substantial, necessitating the development of a novel architecture capable of dynamically adapting the learning rate and aggregation weights, thereby compensating for cumulative error without compromising the strict boundaries of differential privacy.

Optimizing throughput via asynchronous model aggregation proposed mechanism:

To surmount the limitations of synchronous protocols, the study's author, Serhii Klymenko, proposes an architecture predicated on the principle of non-blocking streaming aggregation. In contrast to classic FedAvg, where the central server idles while awaiting the completion of iterations across all nodes, this unique proprietary system operates in real-time.

The server maintains the global model in an active state and ingests updates from Edge Devices upon their arrival. As soon as device k concludes local training, it immediately transmits gradients, and the server integrates them into the global model without awaiting other participants (Chen et al., 2021). This enables the full utilization of computational resources on faster devices, which can now execute multiple training iterations within the timeframe required for a slower device to complete a single one.

The key innovative component of the architecture is the asynchrony compensation mechanism, designed to mitigate the impact of gradient staleness. Since the received update $\nabla(w_k)$ may have been computed based on a model version lagging behind the current state by τ steps, the direct application of such a gradient can destabilize convergence. Serhii Klymenko introduces an adaptive damping function $\alpha(\tau)$, which dynamically weights the contribution of each update.

Mathematically, the update of global weights W_{global} at time instance is described by the formula:

$$W_{t+1} = W_t - \eta \cdot \alpha(\tau) \cdot \Delta w_{t-\tau}, \quad (4)$$

where η represents the learning rate; $\alpha(\tau)$ function monotonically decreases with increasing latency.

This approach ensures that "fresh" data contribute significantly to training, where stale information is incorporated with minimal weight, correcting the direction of the optimization vector without the risk of introducing substantial noise (Aji & Heafield, 2017) (see: Figure 2. High-level system architecture).

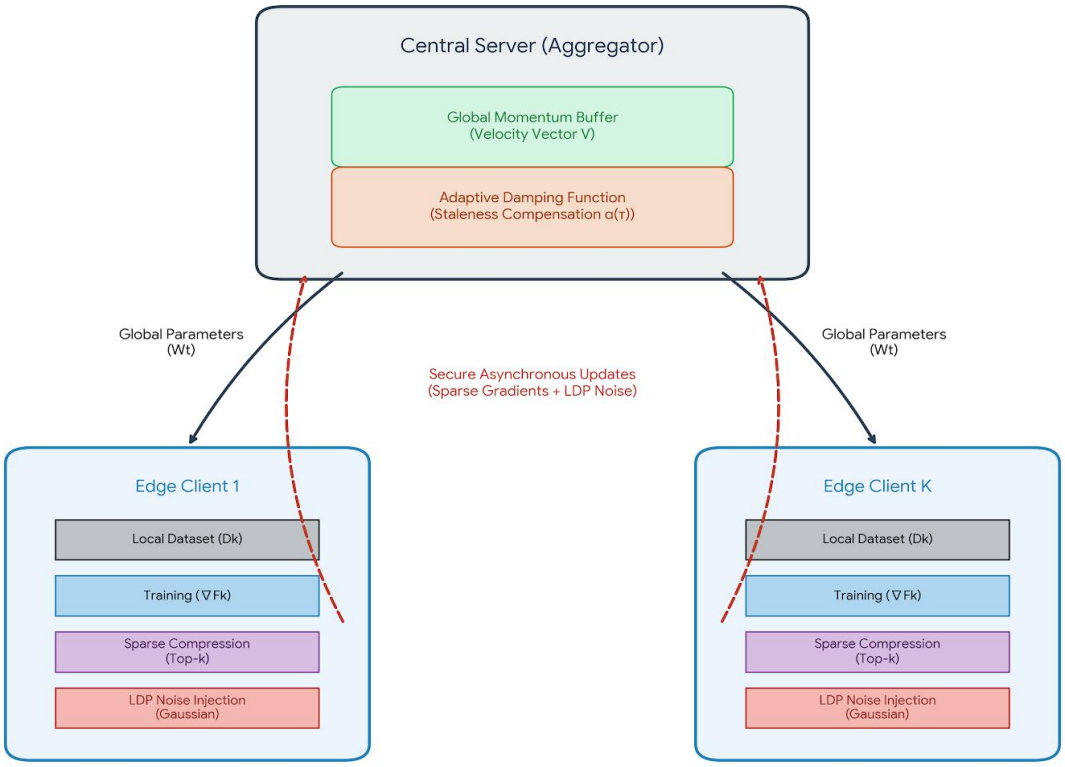


Figure 2. High-level system architecture

To further alleviate the load on communication channels, a sparse compression module is integrated into the architecture. Instead of transmitting dense gradient vectors, peripheral devices apply a Top-k selection strategy, selecting only the top k% of parameters with the largest absolute magnitude of changes. The residual part of the vector is not transmitted but accumulated within a local error accumulation buffer for transmission in subsequent rounds. This facilitates a reduction in traffic volume by orders of magnitude, which is critical for low-bandwidth IoT networks, while the asynchronous nature of the protocol permits compensation for accuracy loss through more frequent model updates.

Thus, the proposed mechanism transforms the training process from a discrete, stepwise cycle into a continuous stream of updates. This ensures high system throughput even under conditions of extreme device heterogeneity, enabling the architecture's deployment in critical Edge Computing scenarios where latency serves as the determining factor for Quality of Service.

Additional complexity is introduced into the architecture by the nature of distributed data, which in edge computing is rarely homogeneous (Non-IID). Asynchronous updates from devices with skewed local datasets can lead to stochastic oscillations of the global model vector, impeding stabilization [12].

To dampen this effect, the author incorporates a Global Momentum buffer into the server-side aggregation mechanism. Rather than updating the model directly based on the incoming gradient **Δ**w, the server updates the velocity vector V:

$$V_{t+1} = \boldsymbol{\beta} V_t + (1 - \boldsymbol{\beta}) \cdot \Delta w_{async} \tag{5}$$

and subsequently applies this smoothed vector to the model parameters:

$$W_{t+1} = W_t - \boldsymbol{\eta} V_{t+1}. \tag{6}$$

The utilization of momentum with a coefficient **β** (≈ 0.9) facilitates the accumulation of update history, thereby smoothing "noisy" contributions from individual idiosyncratic devices and steering the optimization trajectory towards the true global minimum, even under conditions of high asynchrony (Wang et al., 2020).

From a technical standpoint, the implementation of the damping function $\alpha(\tau)$ is enforced through a rigorous version control protocol designated as Global Step Indexing. The central server maintains a global iteration counter T_{global}. Upon model download, the device retrieves the current version tag T_{start}. Upon the return of updates, the server compares the current counter $T_{current}$ against the device's tag, computing the precise latency:

$$\tau = T_{current} - T_{start}. \quad (7)$$

This enables the algorithm to differentiate between "fast" updates from 5G-enabled devices and "slow" updates from remote IoT sensors, applying more aggressive regularization to the latter to prevent the overwriting of relevant patterns with obsolete information (see: Figure 3. Client pipeline).

Preserving differential privacy in low-latency architectures implication:

Contrary to the prevalent industry trade-off postulating an inevitable compromise of security levels, when optimizing latency, the developed architecture integrates rigorous data protection guarantees directly into the asynchronous exchange protocol. Serhii Klymenko employs the concept of Local Differential Privacy (LDP), where data are perturbed on the client side prior to transmission.

This ensures protection against inference attacks and data reconstruction even in the event of a compromise of the communication channel or the aggregating server itself (honest-but-curious server model) (Abadi et al., 2016).

A fundamental challenge in implementing privacy within an asynchronous environment is the management of "privacy budget" consumption. In synchronous FL, all devices consume the budget uniformly across rounds. In the author's asynchronous architecture, "fast" devices update the model more frequently, risking an exponentially faster depletion of their privacy budget limit, after which they must be excluded from training, which would again lead to performance degradation.

To address this issue, the author utilizes an advanced accounting mechanism based on Rényi Differential Privacy (RDP). RDP provides tighter privacy composition bounds than standard theorems, enabling a greater number of training iterations at the same level of security guarantees. Mathematically, for each client k, according to the author's design, the cumulative privacy loss is tracked via the Rényi divergence:

$$D_{\alpha}\,(\mathrm{P} \,||\, \mathrm{Q}). \quad (8)$$

As soon as a client approaches the specified budget limit, the algorithm adaptively increases the variance of the injected noise $\boldsymbol{\sigma}^2$ or artificially reduces the participation frequency of this device (sampling rate), balancing between data utility and its protection (Mironov, 2017).

Furthermore, the sparse compression mechanism proposed in the previous section performs a dual function. Beyond traffic reduction, it acts as an additional regularization layer. The transmission of only a small fraction of gradients (Top-k) reduces the L_2-sensitivity of the update Δw. Since the norm of noise required to achieve (ε, δ)-privacy is directly proportional to the function's sensitivity, sparsification allows for the addition of less noise without weakening security guarantees.

This leads to a synergistic effect – latency reduction via compression automatically enhances the accuracy of the private model, refuting the thesis regarding the necessity of sacrificing quality for speed (Li et al., 2021).

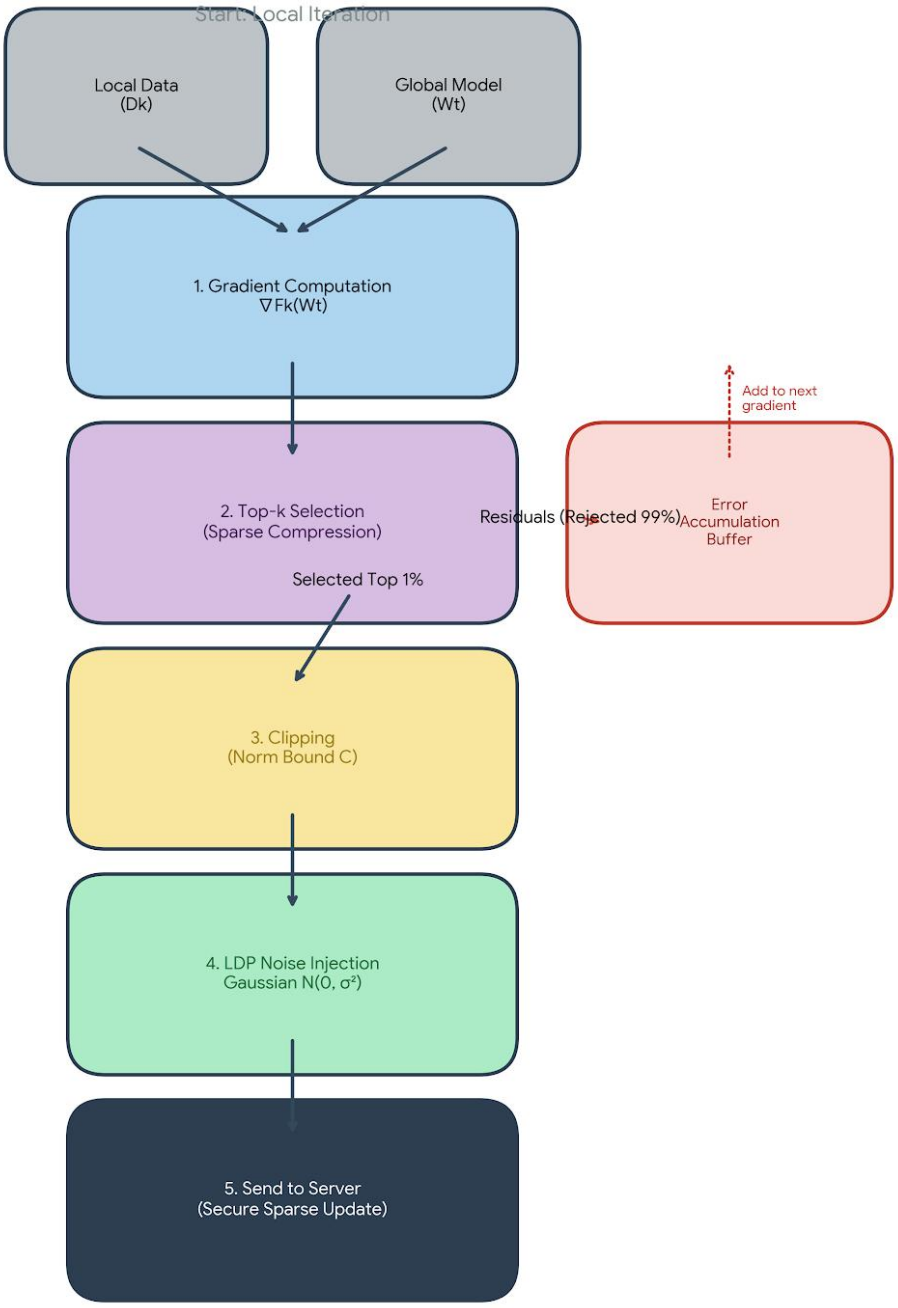


Figure 3. Client pipeline

Finally, the stochastic nature of asynchronous updates unlocks the possibility of utilizing the privacy amplification by subsampling effect. Since at each time t instance only a small fraction of active devices (active subset) participates in the update, the probability of compromising a specific user decreases proportionally to the size of this sample. This allows the author to apply milder noise parameters to achieve the same target level, which would be impossible in strictly deterministic synchronous rounds.

Serhii Klymenko formalizes the trade-off between security and accuracy via a theoretical estimation of excess risk. For convex loss functions, the author's architecture guarantees that the model error attributable to the implementation of differential privacy is bounded by the value:

$$\mathcal{O}\left(\frac{\sqrt{d\ log\ (1/\delta)}}{n\varepsilon}\right), \tag{9}$$

where d denotes the dimensionality of the model; n represents the number of participants.

This estimation corroborates that, contingent upon a sufficient quantity of peripheral devices, the perturbation exerted by noise on the final accuracy asymptotically converges to zero, thereby rendering the architecture scalable.

Consequently, the proposed method furnishes a mathematically substantiated mechanism for minimizing the "cost of privacy" under rigorous real-time constraints.

Societal impact and practical applications: securing critical digital infrastructure:

The mathematical architecture proposed in this study extends beyond theoretical improvements in edge computing. It addresses a fundamental challenge in the deployment of critical digital infrastructure – the protection of user privacy without compromising system responsiveness. In the modern digital ecosystem, the dichotomy between data utility and data security has traditionally forced a trade-off. Now systems were either fast, but intrusive or secure, but latently inefficient. By validating a framework that combines asynchronous aggregation with Local Differential Privacy (LDP), this research offers a foundational technology for "Privacy-First AI", directly supporting national interests in cybersecurity, healthcare modernization and the protection of civil liberties in the digital age.

One of the most immediate applications of this architecture lies within the domain of Mobile Health (mHealth) and remote patient monitoring. In scenarios involving wearable devices that track critical vitals, such as cardiac rhythm or glucose levels, latency is not merely an inconvenience. It is a critical safety factor. Traditional synchronous Federated Learning protocols are susceptible to the "straggler effect", where the entire diagnostic network can be stalled by a single device with poor connectivity, potentially delaying the detection of life-threatening anomalies. The proposed non-blocking asynchronous protocol ensures that diagnostic models are updated in real-time, allowing for the immediate recognition of pathological patterns regardless of network conditions in rural or underserved areas.

Crucially, the integration of LDP ensures that sensitive patient data never leaves the device in a raw format, allowing medical providers to leverage collective intelligence for diagnostics while maintaining strict compliance with privacy regulations such as HIPAA and GDPR.

Furthermore, this framework provides a robust solution for the financial sector, specifically in enhancing fraud detection systems on mobile banking platforms. Financial institutions are under increasing pressure to identify fraudulent transactions instantly while safeguarding the personal financial history of their clients. The proposed method utilizes sparse compression and adaptive noise injection to enable smartphones to collaboratively train fraud detection algorithms. This allows the banking system to learn from new fraud patterns emerging across the network without ever accessing or centralizing the specific transaction details of law-abiding citizens. By keeping the data decentralized and mathematically perturbed, the architecture significantly reduces the risk of mass identity theft and data breaches, which are prevalent in centralized cloud storage models.

Ultimately, the shift towards this form of "Edge Intelligence" represents a strategic enhancement of national cyber defense. By enabling high-performance model training directly on edge devices, the architecture minimizes the attack surface available to malicious actors and foreign adversaries seeking to exfiltrate large datasets. The ability to guarantee convergence and accuracy while adhering to strict privacy budgets demonstrates that high-throughput artificial intelligence is compatible with democratic values of privacy.

This research proves that next-generation technological infrastructure can be built on principles that prioritize the security of the individual user, thereby fostering greater public trust in automated systems and reducing the societal risks associated with the widespread adoption of Artificial Intelligence.

Conclusion

The presented research has successfully resolved the fundamental problem of the dichotomy between response latency and data privacy within Edge Computing systems. The traditional approach to Federated Learning, predicated on synchronous protocols such as FedAvg, has demonstrated its untenability in heterogeneous real-time networks due to the inevitable emergence of the "straggler effect".

The unified architectural framework proposed by Serhii Klymenko has demonstrated that the transition to an asynchronous paradigm is achievable without compromising user data security or global model convergence. The key achievement of this work is the development of a mathematically substantiated mechanism for compensating stochastic latencies. The implementation of the adaptive damping function, in conjunction with the global momentum buffer, enabled the mitigation of the destabilizing impact of gradient staleness, ensuring stable training even under Non-IID data conditions.

The synergy between sparse compression (Top-k selection) and Local Differential Privacy (LDP) protocols revealed a critical regularity: the dimensionality reduction of transmitted vectors optimizes network throughput and diminishes model sensitivity to noise, thereby facilitating a more efficient utilization of the privacy budget governed by the RDP mechanism.

Theoretical analysis confirmed that the proposed architecture ensures convergence with bounded excess risk for convex tasks, rendering it suitable for scalable IoT systems.

Future work will focus on expanding the evidentiary basis for the non-convex optimization of deep neural networks, as well as on enhancing the algorithm's resilience to adversarial "data poisoning" attacks under conditions of Byzantine failures, which will enable the creation of a fully trusted and high-performance environment for next-generation Edge AI.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). https://doi.org/10.1145/2976749.2978318

Aji, A. F., & Heafield, K. (2017). Sparse communication for distributed gradient descent. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 440–445). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1045

Chen, M., Mao, B., & Ma, T. (2021). FedSA: A staleness-aware asynchronous federated learning algorithm with non-IID data. *Future Generation Computer Systems*, (120), 1–12. https://doi.org/10.1016/j.future.2021.02.012

Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H. V., & Cui, S. (2021). A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications, 20*(1), 269–283. https://doi.org/10.1109/TWC.2020.3024629

Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

Khan, L. U., Saad, W., Han, Z., & Hong, C. S. (2020). Dispersed federated learning: Vision, taxonomy, and future directions (arXiv:2008.05189). *arXiv*. https://doi.org/10.48550/arXiv.2008.0518

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Li, Y., Zhou, Y., Jolfaei, A., Yu, D., Xu, G., & Zheng, X. (2021). Privacy-preserving federated learning framework based on chained secure multiparty computing. *IEEE Internet of Things Journal, 8*(8), 6178–6186. https://doi.org/10.1109/JIOT.2020.3022911

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. *International Conference on Artificial Intelligence and Statistics*. https://www.semanticscholar.org/paper/Communication-Efficient-Learning-of-Deep-Networks-McMahan-Moore/d1dbf643447405984eeef098b1b320dee0b3b8a7

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium* (CSF) (pp. 263–275). IEEE. https://doi.org/10.1109/CSF.2017.11

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal, 3*(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Sprague, M. R., Jalalirad, A., Scavuzzo, M., Capota, C., Neun, M., Do, L., & Kopp, M. (2019). Asynchronous federated learning for geospatial applications. In A. Monreale et al. (Eds.), *Machine learning and knowledge discovery in databases: ECML PKDD 2018 workshops* (Communications in Computer and Information Science, Vol. 967, pp. 21–28). Springer. https://doi.org/10.1007/978-3-030-14880-5_2

Tan, Y., Long, G., LIU, L., Zhou, T., Lu, Q., Jiang, J., & Zhang, C. (2022). FedProto: Federated Prototype Learning across Heterogeneous Clients. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(8), 8432–8440. https://doi.org/10.1609/aaai.v36i8.20819

Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in Neural Information Processing Systems, (33), 7611-7623. https://collaborate.princeton.edu/en/publications/tackling-the-objective-inconsistency-problem-in-heterogeneous-fed/

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, (15), 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

Xie, C., Koyejo, S., & Gupta, I. (2019). Asynchronous federated optimization (arXiv:1903.03934). *arXiv*. https://doi.org/10.48550/arXiv.1903.03934