Taming Bandwidth Bottlenecks in Federated Learning via ECN-based Gradient Compression

Javier Palomares*, Chiara Camerota§, Flavio Esposito§, Estefanía Coronado*†,
Cristina Cervelló-Pastor‡, and Muhammad Shuaib Siddiqui*
*i2CAT Foundation, Barcelona, Spain;
Email: {javier.palomares, estefania.coronado, estela.carmona, shuaib.siddiqui}@i2cat.net
†High-Performance Networks and Architectures, Universidad de Castilla-La Mancha, Albacete, Spain;
Email: estefania.coronado@uclm.es
‡Department of Network Engineering, Universitat Politècnica de Catalunya, Castelldefels, Barcelona, Spain;
Email: cristina.cervello@upc.edu
§Department of Computer Science, Saint Louis University, Saint Louis, Missouri, United States of America;
Email: {chiara.camerota, flavio.esposito}@slu.edu

***Abstract*—Communication bottlenecks remain a key challenge in Federated Learning (FL), particularly in dynamic and resource-constrained environments. While compression strategies such as sparsification and quantization reduce communication overhead, they are typically agnostic to runtime variability and the semantic relevance of updates. This paper introduces SCALP (Selective Compression via Adaptive Lightweight Protocol), a novel hybrid communication compression mechanism that jointly considers local gradient variance and uplink bandwidth to guide adaptive filtering decisions. Each worker dynamically selects a compression level mapped to a tunable filtering ratio, balancing communication reduction and update relevance. The selected compression level is encoded as a 2-bit signal embedded in the Explicit Congestion Notification (ECN) field of the IP header, enabling stateless, lightweight signaling without modifying transport-layer protocols. Experimental results on CNN and CNN-LSTM models over the CMAPSS dataset show that SCALP reduces transmitted data by over 25% while maintaining convergence time within 2% and achieving up to 2.15% higher final accuracy compared to baseline methods. Comparative analysis against Deep Gradient Compression (DGC) and bandwidth-aware filtering confirms SCALP's ability to integrate gradient-level relevance and network conditions for robust, efficient training in bandwidth-constrained FL scenarios.**

***Index Terms*—Networks for Learning, Gradient Compression, Bandwidth Adaptation.**

I. Introduction

Network operators and major Internet companies increasingly deploy distributed learning tasks across their infrastructures to optimize internal services such as load balancing, content delivery, and anomaly detection, as well as for user-facing applications like targeted advertisement. Federated Learning (FL) has emerged as a key enabler for training machine learning models without centralized data aggregation, allowing clients to collaboratively train shared models across heterogeneous nodes. However, efficiently executing FL on the edge-to-cloud continuum presents significant challenges due to dynamic network conditions, resource heterogeneity, congestion, and limited coordination [1]–[4].

In such environments, communication bottlenecks often prolong training convergence, resulting in increased energy consumption and latency [5], [6]. Conventional FL architectures typically assume static placement of parameter servers (PS), which exacerbates inefficiencies when network variability affects communication links. To solve these issues, prior work has proposed model compression techniques such as Deep Gradient Compression (DGC) [7] and FedZip [8], which reduce uplink traffic through sparsification and quantization. However, these strategies typically apply static compression policies, making them agnostic to runtime network conditions.

In heterogeneous edge deployments, static compression can lead to information loss under degraded links, exacerbate straggler effects when clients operate with asymmetric bandwidth, and increase convergence time when multiple jobs compete for limited resources [9], [10]. Moreover, existing techniques often treat all gradient updates equally, regardless of their contribution to model convergence, resulting in inefficient bandwidth usage. Recent advances in semantic communication have emphasized the importance of transmitting task-relevant information over raw gradients. By extracting semantically meaningful components, these approaches aim to retain only the most relevant updates [11], [12]. Adaptive compression strategies have also emerged to improve communication efficiency and fairness. For instance, AdaGQ [13] adjusts gradient quantization based on local variance and device capabilities, while Caesar [14] employs data-aware heuristics to improve update quality under non-IID conditions. FedCG [15] jointly optimizes client selection and compression to mitigate straggler effects, and AdapComFL [16] predicts uplink bandwidth capacity to dynamically adjust sketch sizes. While these approaches advance the state of adaptive compression, they remain largely agnostic to the semantic relevance of updates and their direct impact on model convergence.

To address these challenges, SCALP (Selective Compression via Adaptive Lightweight Protocol) is introduced as an adaptive gradient compression mechanism that improves communication efficiency and training robustness in FL. SCALP employs a hybrid decision policy that enables each worker to select a compression level based on both the statistical variance of

its local gradient updates and the current uplink bandwidth condition. This strategy allows the system to prioritize the transmission of semantically relevant updates under constrained network conditions, reducing communication overhead without compromising learning performance. The selected compression level is encoded using a 2-bit signal embedded into the Explicit Congestion Notification (ECN) field of the IP header, supporting lightweight, stateless coordination with the PS. While ECN-based signaling primarily targets intra-domain deployments, such as edge-cloud infrastructures or managed data center environments, the adaptive compression policy of SCALP, which is grounded in gradient variance and bandwidth awareness, is a generalizable strategy. It could be adapted with alternative signaling mechanisms to support broader applicability in wide-area FL scenarios.

II. System Design

To improve communication efficiency and training robustness under variable network conditions, SCALP introduces an adaptive compression mechanism that reacts to both learning dynamics and real-time bandwidth observations. In contrast to static schemes that map bandwidth quartiles to fixed compression levels, SCALP jointly considers the statistical relevance of local updates and current link quality to inform compression decisions.

Each worker $n \in \mathcal{N}$ begins by computing the variance of its local gradient vector, used as a proxy for the diversity and magnitude of its update. This variance, defined in (1), is computed over the gradient components g_i with respect to their mean μ. A low variance indicates minimal deviation from the current global model, suggesting limited contribution, while high variance reflects a more substantial update likely to influence convergence.

$$\sigma_g^2 = \frac{1}{|I|} \sum_{i \in I} (g_i - \mu)^2, \tag{1}$$

Simultaneously, the node monitors its uplink bandwidth B_n through passive throughput sampling and telemetry. As defined in (2), the compression level $C_n \in \{0, 1, 2, 3\}$ is selected using a rule-based policy that combines gradient variance σ_g^2 and low-bandwidth threshold T_{low}.

$$C_n = \begin{cases} 0, & \text{if } \sigma_g^2 < \theta \text{ and } B_n < T_{\text{low}}, \\ 1, & \text{if } \sigma_g^2 < \theta \text{ and } B_n \geq T_{\text{low}}, \\ 2, & \text{if } \sigma_g^2 \geq \theta \text{ and } B_n < T_{\text{low}}, \\ 3, & \text{if } \sigma_g^2 \geq \theta \text{ and } B_n \geq T_{\text{low}}. \end{cases} \tag{2}$$

The variance threshold is set to $\theta = 10^{-3}$, based on empirical observations during model convergence. This value aligns with findings in [17], which show that gradient variance typically decreases as training progresses and serves as a reliable indicator of low-impact updates. The low-bandwidth threshold is defined as $T_{\text{low}} = 5$ Mbps, reflecting congestion levels commonly observed in edge deployments. Prior analyses indicate that uplink throughput in such environments, particularly under contention or near cell boundaries, often ranges between 5 and 10 Mbps [18]. Selecting $T_{\text{low}} = 5$ Mbps ensures that compression is selectively applied under constrained conditions, while remaining representative of realistic intra-domain scenarios. The impact of varying both θ and T_{low} on training performance and communication efficiency is further evaluated in Section III.

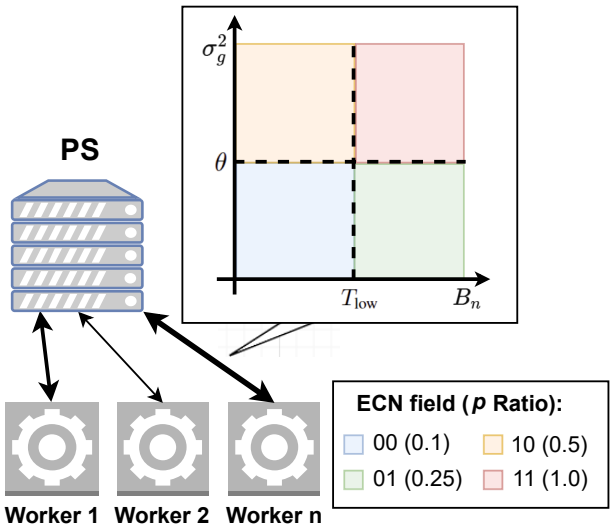

Fig. 1. SCALP overview: Each worker independently selects a filtering ratio based on local gradient variance and observed uplink bandwidth. Line width reflects link capacity, and the PS decodes ECN values to interpret each worker's filtering level.

Moreover, the selected compression level C_n is mapped to a filtering ratio $p \in \{0.1, 0.25, 0.5, 1.0\}$, where lower values of C_n correspond to more aggressive filtering, thereby reducing the communication payload. These specific ratios are informed by established sparsification studies in adaptive gradient compression [7], [19], [20], which investigate the effect of retaining various proportions of gradient components. These works demonstrate that retention fractions as low as 10% can still preserve convergence under certain conditions, while enabling substantial communication savings. Although this exact progression is not standardized, it provides a balanced range from highly constrained to unconstrained settings, supporting flexible adaptation to bandwidth availability while preserving gradient significance. The lowest ratio ($p = 0.1$) aggressively suppresses low-impact updates under limited connectivity, whereas the highest ratio ($p = 1.0$) retains the full gradient to preserve fidelity in favorable conditions. Intermediate values are selected to offer practical granularity without incurring the complexity of finer-scale tuning. This design simplifies runtime decision-making and supports efficient compression control with minimal signaling overhead.

To communicate the compression level to the PS without coordination overhead, each worker encodes the 2-bit value of C_n into the ECN field of the IP header [21]. This signaling approach eliminates the need for protocol changes or explicit control messages and remains compatible with most deployment scenarios, such as edge clusters and data centers. While ECN may be stripped by legacy or wide-area

middleboxes, prior studies [22] show that less than 1% of ECN-capable paths fail due to such interference. SCALP targets controlled environments where ECN semantics are preserved. Fallback mechanisms can be used if ECN support is unavailable, but alternative headers like DSCP [23] or TCP options [24] involve higher protocol overhead or limited support, making ECN a practical and efficient solution.

As illustrated in Fig. 1, each worker dynamically selects a compression level based on its local gradient variance and uplink bandwidth conditions. This level is encoded as a 2-bit value embedded in the ECN field of the IP header and transmitted with the gradient updates. The line width in the figure reflects the available link capacity. On the receiver side, the PS decodes the ECN field to infer the filtering ratio applied by each worker, enabling bandwidth-aware and variance-sensitive aggregation. This stateless signaling mechanism avoids explicit coordination or protocol modifications and supports consistent model convergence under heterogeneous network conditions.

III. Evaluation

This section presents a comprehensive evaluation of SCALP under realistic FL conditions using the CMAPSS dataset [25] and the hybrid Convolutional Neural Network (CNN) Long Short-Term Memory (LSTM) architecture proposed in [26]. The evaluation is structured in three phases. First, SCALP is evaluated against the training baseline from [26] to assess its impact on training time, communication overhead, and model accuracy. The default thresholds used throughout the evaluation are $\theta = 10^{-3}$ and $T_{\text{low}} = 5\,\text{Mbps}$, selected based on observed convergence behavior [17] and common congestion levels in edge networks [18], respectively. Next, a parameter impact analysis quantifies how variations in SCALP's two compression parameters, the gradient variance threshold θ and the low-bandwidth threshold T_{low}, influence performance trade-offs. Finally, SCALP is benchmarked against two representative compression strategies: bandwidth-aware filtering [27] and DGC [7], to validate the advantages of its dual-adaptive design. The accuracy target is set to 90% across all training iterations to enable fair convergence comparisons. To ensure statistical robustness, each performance metric is averaged over ten thousand independent cycles and reported with a 95% confidence interval.

Table I compares SCALP with the baseline [26] using the default thresholds $\theta = 10^{-3}$ and $T_{\text{low}} = 5\,\text{Mbps}$. SCALP reduces total transmitted data by approximately 24.9% for CNN and 25.8% for CNN-LSTM, while maintaining convergence times within 2.0% of the baseline. These improvements are achieved through bandwidth-aware filtering, which prioritizes the transmission of semantically relevant gradients.

Due to its sequence modeling capabilities and favorable trade-off between communication efficiency and training performance, the CNN-LSTM model is selected to evaluate SCALP under deployment-specific configurations. The analysis examines the two main compression triggers: the gradient variance threshold θ and the low-bandwidth threshold T_{low}.

TABLE I
Comparison of Training Time and Transmitted Data between SCALP and the Baseline [26] for CNN and CNN-LSTM Models.

Metric	Baseline [26]	SCALP
CNN Model		
Training Time to 90% Accuracy (s)	2165.91 ± 31.96	2210.02 ± 21.44
Total Data Transmitted (KB)	498.66 ± 4.28	374.75 ± 2.80
CNN-LSTM Model		
Training Time to 90% Accuracy (s)	4544.23 ± 33.27	4579.66 ± 22.14
Total Data Transmitted (KB)	394.44 ± 1.81	292.76 ± 1.62

TABLE II
Evaluation Scenarios for SCALP Threshold Parameter Impact.

Scenario	Parameter	Metrics	Values
1	θ	- Final Accuracy - Transmitted Data	$10^{-4}, 10^{-3}, 10^{-2}$
2	T_{low}	- Training Time - Transmitted Data	2, 5, 10 Mbps

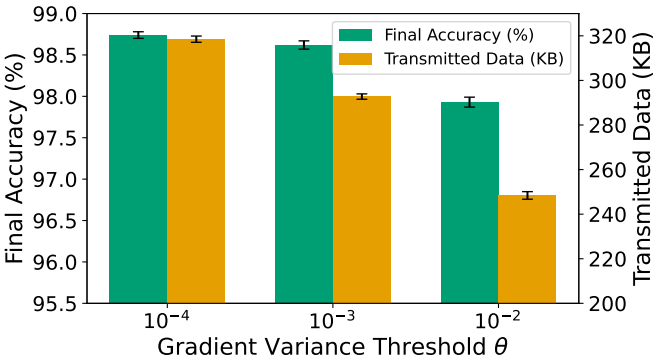

Fig. 2. Effect of the Gradient Variance Threshold θ on SCALP Transmitted Data and Model Final Accuracy.

Both parameters are selected based on empirical convergence behavior and prior studies [17], [18]. Their influence on training dynamics and communication overhead is quantified, with the evaluation scenarios, corresponding metrics, and parameter ranges summarized in Table II.

Figure 2 analyzes the impact of the gradient variance threshold θ on SCALP's communication efficiency and model accuracy. This parameter determines how each worker assesses the statistical significance of its local gradient update. Specifically, when the variance of the gradient vector falls below θ, the update is considered to have low impact and is subject to more aggressive filtering. As θ increases, the condition $\sigma_g^2 < \theta$, defined in (1), is satisfied more frequently, resulting in a larger fraction of updates being classified as insignificant and filtered locally instead of being transmitted. This behavior leads to a significant reduction in communication volume, with over 20% difference observed between the lowest and highest threshold values. However, this gain in bandwidth efficiency involves a trade-off. At the most aggressive setting ($\theta = 10^{-2}$),

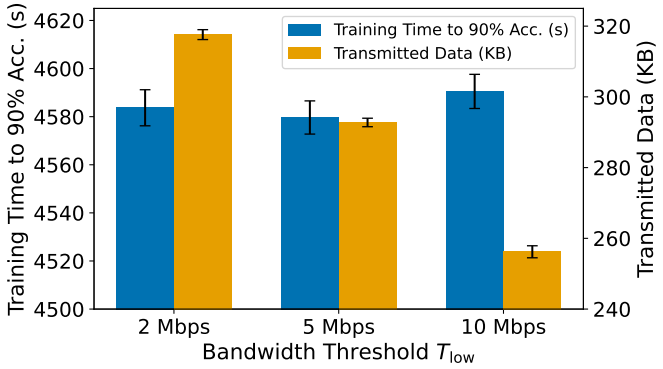

Fig. 3. Effect of the Low-Bandwidth Threshold T_{low} on SCALP's Training Time to Reach 90% Accuracy and Transmitted Data.

compression removes gradient components that, despite their low variance, still contribute to model convergence. This effect results in a modest decrease in final accuracy.

Figure 3 illustrates the effect of the low-bandwidth threshold T_{low} on SCALP's training time to reach 90% accuracy and the total data transmitted. This threshold does not represent the actual measured bandwidth, but rather serves as a decision criterion used by each worker to determine whether the current uplink condition should be treated as congested. When the observed bandwidth B_n falls below this threshold, compression is triggered. Lower threshold values, such as 2 Mbps, imply that only severely degraded links activate compression, leading to more frequent full-gradient transmissions and higher overall data exchange. In contrast, higher thresholds, such as 10 Mbps, cause workers to classify even moderately loaded links as congested. Consequently, compression is applied more aggressively, resulting in a substantial reduction in transmitted data. However, this increased compression sensitivity can impact training dynamics: excessive filtering may discard semantically relevant updates, resulting in delayed convergence. This effect is reflected in a slight increase in training time at the highest threshold setting.

These results indicate that θ and T_{low} provide tunable control over SCALP's behavior. Lower values prioritize fidelity by limiting compression, whereas higher values enhance communication efficiency. In both cases, mid-range configurations, specifically a variance threshold of $\theta = 10^{-3}$ and a low-bandwidth threshold of $T_{\text{low}} = 5$ Mbps, are used as robust and empirically validated default values. Therefore, these values are adopted in the subsequent evaluations to ensure a balanced trade-off between accuracy and communication cost.

The final stage of the evaluation compares SCALP against two representative compression strategies. The first is a bandwidth-aware method [27] that filters gradients based exclusively on observed uplink capacity. The second is DGC [7], which retains only gradient components exceeding a global magnitude threshold. These baselines provide a contrast to SCALP's dual-adaptive approach, which combines gradient

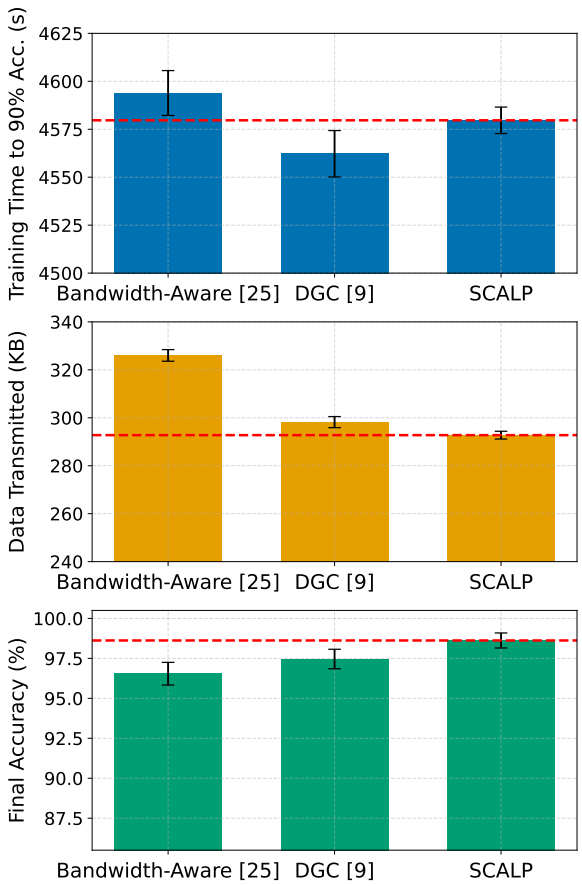

Fig. 4. Performance comparison between SCALP, Bandwidth-Aware Filtering [27], and DGC [7]. The top subplot reports the training time to reach 90% accuracy, the middle subplot shows the total data transmitted, and the bottom subplot presents the final model accuracy. All metrics are measured using a CNN-LSTM model trained on the CMAPSS dataset [25].

variance and bandwidth measurements to make more informed compression decisions.

Figure 4 reports the comparative performance of SCALP against the baseline methods. The top subplot shows the training time required to reach the target accuracy of 90%. All three methods achieve comparable convergence speeds, with SCALP completing training in 4579.7 seconds, which is 0.38% slower than DGC and 0.31% faster than the bandwidth-aware approach. These results indicate that SCALP's compression dynamics do not introduce delays in convergence. The middle subplot illustrates the total data volume exchanged during training. SCALP achieves the highest communication efficiency, reducing data transmission by 10.2% compared to the bandwidth-aware method and by 1.8% relative to DGC. These savings reflect the advantage of integrating both variance-driven filtering and bandwidth awareness. The bottom subplot presents the final model accuracy. SCALP achieves 98.62%, outperforming DGC by 1.19% and the bandwidth-aware strategy by 2.15%. This improvement underscores the limitations of compression strategies that rely exclusively on gradient magnitude or network state, as they may discard updates that are statistically relevant to model convergence.

IV. Conclusions

This paper presented SCALP, a lightweight, network-aware gradient compression mechanism for Federated Learning that combines adaptive filtering with stateless signaling via the ECN field. By allowing each client to adjust its compression level based on local gradient variance and uplink bandwidth conditions, SCALP reduces communication overhead without compromising convergence accuracy. Evaluations on the CMAPSS dataset using CNN and CNN-LSTM models show that SCALP consistently outperforms baseline strategies in both efficiency and model quality. Its dual-adaptive design, which integrates statistical relevance and link quality, proves effective in maintaining performance under constrained network conditions. These results establish SCALP as a practical and robust solution for communication-efficient FL in dynamic and heterogeneous edge environments.

Acknowledgment

This work was supported by NSF awards 2201536 and 2430236. The work of J. Palomares was conducted at SLU, supported by MCIN/AEI/10.13039/501100011033 (FEDER "a way of making Europe") under grant PID2022-142332OA-I00.

References

[1] C. Li, X. Zeng *et al.*, "PyramidFL: a fine-grained client selection framework for efficient federated learning," *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, p. 158–171, 2022. [Online]. Available: https://doi-org.recursos.biblioteca.upc.edu/10.1145/3495243.3517017

[2] W. Gao, O. Tavallaie *et al.*, "Federated Learning as a Service for Hierarchical Edge Networks with Heterogeneous Models," *Service-Oriented Computing*, pp. 85–99, 2025.

[3] S. Kalra, J. Wen *et al.*, "Decentralized Federated Learning through Proxy Model Sharing," *Nature Communications*, vol. 14, no. 1, may 2023. [Online]. Available: http://dx.doi.org/10.1038/s41467-023-38569-4

[4] Z. Zhang, A. Pinto *et al.*, "Privacy and Efficiency of Communications in Federated Split Learning," *IEEE Transactions on Big Data*, pp. 1–12, May 2023.

[5] J. Li, H. Xu *et al.*, "Lyra: Elastic Scheduling for Deep Learning Clusters," *Proceedings of the Eighteenth European Conference on Computer Systems*, p. 835–850, 2023. [Online]. Available: https://doi-org.recursos.biblioteca.upc.edu/10.1145/3552326.3587445

[6] J. Liu, F. Lai *et al.*, "Venn: Resource Management for Collaborative Learning Jobs," 2025. [Online]. Available: https://arxiv.org/abs/2312.08298

[7] Y. Lin, S. Han *et al.*, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," 2020. [Online]. Available: https://arxiv.org/abs/1712.01887

[8] A. Malekijoo, M. J. Fadaeieslam *et al.*, "FEDZIP: A Compression Framework for Communication-Efficient Federated Learning," 2021. [Online]. Available: https://arxiv.org/abs/2102.01593

[9] A. Kumar and S. N. Srirama, "FedStrag: Straggler-aware federated learning for low resource devices," *Digital Communications and Networks*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235286482400169X

[10] A. Reisizadeh, I. Tziotis *et al.*, "Straggler-Resilient Federated Learning: Leveraging the Interplay Between Statistical Accuracy and System Heterogeneity," *IEEE Journal on Selected Areas in Information Theory*, vol. PP, pp. 1–1, 06 2022.

[11] Y. Bian, P. Liu *et al.*, "Federated Learning and Semantic Communication for the Metaverse: Challenges and Potential Solutions," *Electronics*, vol. 14, no. 5, p. 868, 2025.

[12] S. D. Okegbile, H. Gao *et al.*, "FLeS: A Federated Learning-Enhanced Semantic Communication Framework for Mobile AIGC-Driven Human Digital Twins," *TechRxiv*, 2024, preprint.

[13] H. Liu, F. He *et al.*, "Communication-Efficient Federated Learning for Heterogeneous Edge Devices Based on Adaptive Gradient Quantization," 2022. [Online]. Available: https://arxiv.org/abs/2212.08272

[14] J. Yan, J. Liu *et al.*, "Caesar: A Low-deviation Compression Approach for Efficient Federated Learning," 2024. [Online]. Available: https://arxiv.org/abs/2412.19989

[15] Y. Wu, Y. Kang *et al.*, "FedCG: Leverage Conditional GAN for Protecting Privacy and Maintaining Competitive Performance in Federated Learning," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, ser. IJCAI-2022. International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, p. 2334–2340.

[16] Y. Zhuansun, D. Li *et al.*, "Communication-Efficient Federated Learning with Adaptive Compression under Dynamic Bandwidth," 2024. [Online]. Available: https://arxiv.org/abs/2405.03248

[17] F. Faghri, D. Duvenaud *et al.*, "A Study of Gradient Variance in Deep Learning," 2020. [Online]. Available: https://arxiv.org/abs/2007.04532

[18] Ericsson, "Optimizing indoor connectivity – Mobility Report," Online, Nov. 2021, [Accessed 05-07-2025]. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/mobile-broadband-indoor-deployment

[19] C. Renggli, S. Ashkboos *et al.*, "SparCML: High-Performance Sparse Communication for Machine Learning," 2019. [Online]. Available: https://arxiv.org/abs/1802.08021

[20] S. Stich, J.-B. Cordonnier *et al.*, "Sparsified SGD with Memory," 2018. [Online]. Available: https://arxiv.org/abs/1809.07599

[21] J. Postel, "Internet Protocol," Request for Comments 791, 1981, [Online] Accessed 05-07-2025. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3168

[22] M. Kühlewind, S. Neuner *et al.*, "On the State of ECN and TCP Options on the Internet," in *Passive and Active Measurement (PAM)*, M. Roughan and R. Chang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 135–144.

[23] K. Nichols, S. Blake *et al.*, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," Request for Comments 2474, 1998, [Online] Accessed 05-07-2025. [Online]. Available: https://www.rfc-editor.org/rfc/rfc2474

[24] V. Jacobson, R. Braden *et al.*, "TCP Extensions for High Performance," Request for Comments 1323, 1992, [Online] Accessed 05-07-2025. [Online]. Available: https://www.rfc-editor.org/rfc/rfc1323

[25] D. Inc., "Predictive Maintenance." [Online]. Available: https://github.com/kokikwbt/predictive-maintenance/tree/main

[26] A. Wahid, J. G. Breslin *et al.*, "Prediction of Machine Failure in Industry 4.0: A Hybrid CNN-LSTM Framework," *Applied Sciences*, vol. 12, no. 9, 2022.

[27] Z. Tang, J. Huang *et al.*, "Bandwidth-Aware and Overlap-Weighted Compression for Communication-Efficient Federated Learning," in *Proceedings of the 53rd International Conference on Parallel Processing*. ACM, Aug. 2024, p. 866–875. [Online]. Available: http://dx.doi.org/10.1145/3673038.3673142