

Communication-Efficient Personalized Federated Learning: A Comprehensive Investigation of Strategies and Challenges

Shuyu Kang

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

skangap@connect.ust.hk

Abstract. Personalized Federated Learning (PFL) has emerged as a key paradigm to address statistical heterogeneity in distributed machine learning, tailoring models to individual client data. However, personalization strategies often exacerbate the communication bottleneck, a critical challenge compounded by system heterogeneity across edge devices. This survey provides a comprehensive and structured overview of communication optimization techniques within PFL. This paper proposes a novel taxonomy that categorizes the state-of-the-art into three pillars: (1) architectural methods, such as parameter-efficient learning and structured pruning, that minimize the size of model updates; (2) adaptive process methods, including personalized compression and resource-aware client selection, that optimize the communication protocol; (3) topological frameworks, like clustered and decentralized learning, that restructure collaborative pathways. For each category, this paper analyzes the intricate trade-offs between personalization effectiveness, communication cost, and system complexity. Furthermore, this paper highlights critical challenges related to fairness, robustness, and security that arise from these optimization techniques. Finally, this paper identifies open challenges and chart future research directions, advocating for hybrid, automated frameworks that co-design for efficiency, equity, and security to advance the deployment of robust PFL systems in the real world.

Keywords: Personalized Federated Learning (PFL), Communication Efficiency, System Heterogeneity, Statistical Heterogeneity.

1. Introduction

The proliferation of data-driven models is fueled by vast datasets, yet a significant portion of this data is generated on privacy-sensitive edge devices. Traditional centralized training, which requires uploading raw data to a server, faces critical challenges regarding privacy, communication costs, and data silos [1]. Federated Learning (FL) was introduced as a decentralized paradigm to overcome these barriers [2]. In FL, clients collaboratively train a global model by only sharing model updates with a central server, while their raw data remains localized. This approach, exemplified by the seminal Federated Averaging (FedAvg) algorithm, fundamentally protects user privacy. Despite its promise, FL's real-world deployment is hindered by several key challenges [1]. These include statistical heterogeneity, arising from Non-Independent and Identically Distributed (Non-IID) data across clients [2]; system heterogeneity, due to variances in client hardware, network conditions, and availability; and the persistent communication bottleneck, which often constrains the overall training efficiency. These intertwined challenges necessitate novel optimization strategies beyond conventional distributed learning.

The "one-model-fits-all" approach of conventional FL is often suboptimal in the face of statistical heterogeneity, as a single global model may fail to generalize well to the diverse data distributions of individual clients [3, 4]. To address this, Personalized Federated Learning (PFL) has emerged as a prominent research direction. PFL aims to move beyond a single global model by creating customized models tailored to each client's unique data, thereby bridging the gap between collaborative training and individual performance needs [4].

An extensive body of work has proposed various PFL strategies, including methods based on fine-tuning, multi-task learning, and meta-learning [3, 4]. However, the primary focus of these techniques has been on tackling the challenges of statistical heterogeneity. Less attention has been paid to the

significant communication overhead that these personalization strategies often introduce or exacerbate. The need to train and manage numerous personalized models can increase the complexity and volume of information exchanged. This communication bottleneck becomes even more critical when compounded by system heterogeneity, where clients with limited bandwidth or computational power can severely hinder the training process. Consequently, a knowledge gap exists in the systematic understanding and optimization of communication efficiency within PFL, especially under the dual constraints of data and system heterogeneity [5].

Motivated by this identified gap, this survey presents a comprehensive and structured overview of communication optimization within Personalized Federated Learning, with a specific focus on methods addressing the dual challenges of data and system heterogeneity. The objective of this survey is to establish a clear taxonomy of existing approaches, analyze their inherent trade-offs, and illuminate the intricate relationship between personalization effectiveness and communication cost. This survey makes three primary contributions: (1) it proposes a novel taxonomy to systematically categorize the state-of-the-art; (2) it provides a critical analysis of current methods, highlighting their strengths and limitations; and (3) it identifies key open challenges and suggest promising future research directions. By synthesizing the current landscape, this work aims to provide a valuable roadmap for researchers and practitioners seeking to develop more efficient and robust PFL systems in complex, real-world environments.

2. Taxonomy of Communication Optimization Strategies

2.1. Structured Personalization and Parameter-Efficient Learning

This category addresses the communication bottleneck by modifying the model architecture itself. These methods restructure the model to separate shared, generalizable parameters from personalized, client-specific ones. This design ensures only a small subset of parameters is transmitted, achieving inherent communication efficiency.

2.1.1 Explicit Layer Separation

The explicit separation of shared and personalized layers is a straightforward strategy. This classic "split-model" approach typically partitions a model into a shared backbone, which is collaboratively trained and aggregated across all clients, and a private head, which is owned by each client and never uploaded to the server. By pre-defining the communication boundary at the architectural level, this method drastically reduces the data payload in each round, as only the parameters of the shared layers participate in the federated process. A more sophisticated implementation of this idea is found in HeurpFedLA, which employs a hypernetwork to dynamically identify the layers that are most critical for each client's personalization. These identified layers are then retained locally and excluded from the aggregation and transmission process, enabling a data-driven and dynamic form of layer separation that optimizes the trade-off between personalization and communication cost [6].

2.1.2 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning, inspired by advancements in large language models (LLMs), is another highly effective approach. This paradigm freezes most of a large pre-trained model's parameters and introduces only a few new, trainable modules for adaptation. In personalized federated learning, PEFT offers a dramatic communication advantage. For instance, in a federated fine-tuning framework that leverages Low-Rank Adaptation (LoRA), clients do not exchange updates for the entire massive model. Instead, they only need to communicate the updates for the newly introduced, lightweight low-rank matrices, which have far fewer parameters [7]. This technique can reduce communication overhead by orders of magnitude, making large model personalization feasible on edge devices with limited bandwidth.

2.1.3 Personalized Pruning and Sparsification

Personalized model pruning and sparsification also achieve communication efficiency by having each client learn a unique and sparse sub-network from a larger, denser global model. A canonical example of this is the Federated Modular Network (FedMN). This framework maintains a shared "pool of modules" on the server, and each client utilizes a lightweight routing hypernetwork to selectively activate a subset of these modules, effectively assembling a custom sparse architecture tailored to its local data. This mechanism is communication-efficient because clients only need to upload and download the parameters of their specific activated modules, rather than the entire module pool. This significantly reduces the size of the transmitted data, as the personalized sub-network is considerably smaller than the comprehensive super-network from which it is derived [8].

2.2. Adaptive Communication and Participant Scheduling

Unlike methods that modify the model structure, this category optimizes the communication process itself. The core philosophy is to leverage inherent client heterogeneity instead of viewing it as an impediment. Rather than applying a uniform policy, these approaches dynamically tailor the communication behavior and participation for each client. The strategies are adaptive to the real-time state of the federated network, aiming to improve the efficiency of each communication or reduce the rounds needed for convergence.

2.2.1 Personalized Communication Compression

This work moves away from a single, fixed-rate compression technique for all clients, as a uniform level is suboptimal in a heterogeneous environment. Clients with better network connections can send high-fidelity updates, while those on slower networks should use more aggressive compression to ensure timely delivery and save resources. Personalization is often guided by the client's system state or the statistical importance of its update. For instance, the DAdaQuant framework introduces a client-adaptive quantization scheme where the quantization level for each client is determined by its contribution weight to the global model aggregation. By allocating a larger communication budget to more influential clients, this method ensures that critical updates are preserved with greater fidelity, while still aggressively compressing less significant updates to achieve overall communication savings [9].

2.2.2 Adaptive Local Computation

This idea involves strategically trading local computation for communication rounds. Standard Federated Averaging prescribes a fixed number of local epochs, which fails to account for system heterogeneity and leads to a "straggler" problem that bottlenecks the entire system. Adaptive local computation methods break this rigid structure by letting more capable clients perform more local updates before communicating. By performing more intensive local training, a client can produce a more refined model update, potentially reducing the total communication rounds needed to reach a target accuracy. A method for dynamic epoch adjustment formalizes this trade-off by establishing a theoretical correlation between the number of local epochs and the required communication rounds, allowing the system to dynamically determine the optimal number of local training steps to minimize overall communication frequency [10].

2.2.3 Contribution and Resource-Aware Client Selection

This strategy refines client selection, moving from random sampling to a utility-driven approach. The core insight is that each communication round is a valuable resource that should be allocated to clients offering the best combination of high-value data and system resources. Instead of random selection, these methods evaluate each client based on a composite utility function. This function typically balances a client's statistical utility (the potential contribution of its local data) with its system utility (reliability and efficiency). As systematically reviewed in recent surveys, the overarching goal is to prioritize clients that can provide not only informative updates but can also

transmit them quickly and reliably, thereby accelerating model convergence and reducing the total number of communication rounds needed to complete the training process [11].

2.3. Personalized Communication Topologies and Collaborative Frameworks

This category departs from the conventional federated learning paradigm by restructuring the communication architecture. Instead of a fixed, server-centric "star" topology, these approaches build flexible, personalized collaborative relationships. The objective is to shift from a one-to-all model to a targeted structure, allowing client interactions to be more aligned with their data distributions, task goals, or proximity. This architectural shift reduces redundant global communication and directs information flow where most beneficial, enhancing overall efficiency.

2.3.1 Clustered Federated Learning

Clustered federated learning partitions clients into groups with similar data distributions, based on the premise that specialized cluster models are more effective than a single global one for a diverse population. Communication is personalized at a group level, with model aggregation occurring mainly within each cluster. This significantly reduces communication overhead by preventing clients with dissimilar data from exchanging counterproductive updates. For example, the Iterative Clustered Federated Learning (ICFL) framework dynamically discovers the optimal number of clusters by evaluating the geometric relevance between client models. In this system, clients primarily contribute to and receive updates for their specific cluster model, making each communication round more relevant and efficient by reducing the scope of aggregation from a global to a cluster level [12].

2.3.2 Federated Learning via Knowledge Distillation

This approach personalizes communication by changing its content through knowledge distillation (KD). Instead of transmitting large model parameters, clients exchange compact representations of "knowledge." In the federated context, clients' local models act as "teachers." They process a shared, unlabeled public dataset and upload their output logits, which the server aggregates to train a global "student" model. As a comprehensive survey on the topic explains, this method dramatically reduces communication costs because the knowledge carrier (the logit vector) is significantly smaller in size than the full set of model parameters. This shift from parameter sharing to knowledge sharing not only slashes the data volume per transmission but is also inherently robust to model heterogeneity, as the exchanged knowledge is architecture-agnostic [13].

2.3.3 Decentralized and Device-to-Device Collaboration

As a radical departure from the standard model, decentralized federated learning (DFL) eliminates the central server, with clients communicating directly in a peer-to-peer (P2P) fashion. The communication topology can be static or dynamic. This approach personalizes the communication pathways, as each client's partners are defined by the network structure, not a central authority. As detailed in surveys of the DFL landscape, this server-less architecture provides a primary communication advantage by removing the central server as a potential bottleneck. It reduces overall network latency and resource consumption by localizing communication and obviating the need for every client to establish a connection with a single, potentially distant, coordinating entity, making it a highly efficient model for truly edge-native collaborative learning [14].

3. Discussion

3.1. Challenges

Although promising, the outlined communication optimization strategies face practical challenges and trade-offs. Gains in communication efficiency often come at the cost of increased complexity, potential performance degradation, or new systemic vulnerabilities. Acknowledging these limitations is crucial for developing more robust and holistic solutions.

3.1.1 Personalization Effectiveness vs. Communication Overhead

A fundamental trade-off exists between personalization granularity and communication efficiency. Methods in the first category, such as PEFT and structured pruning, achieve remarkable communication savings by severely restricting the number of parameters that are trained and exchanged. This, however, raises a critical question: whether the limited capacity of these sparse modules is sufficient to capture unique client data characteristics. Over-emphasizing parameter compression can risk undermining the model's expressive power and, consequently, the effectiveness of the personalization. This tension is also evident in methods like knowledge distillation. While transmitting compact logit vectors instead of full model parameters is highly efficient, it introduces the risk of information loss during the distillation process, where nuances from the local "teacher" models may not be fully captured by the global "student" model, potentially leading to a degradation in accuracy [13]. Similarly, retaining a larger number of personalized layers locally can significantly reduce communication costs but may also compromise the global model's performance by limiting its exposure to diverse client knowledge [6].

3.1.2 Complexity and System Overhead of Adaptive Strategies

Another major challenge is the complexity and system overhead introduced by adaptive strategies. The intelligent, dynamic methods from the second category, while powerful in principle, are not without cost. Strategies like personalized compression and contribution-aware client selection depend on having real-time, fine-grained metadata about each client's network conditions, computational power, and data distribution. Acquiring this metadata can itself introduce significant overhead and may raise new privacy concerns. For instance, the DAdaQuant framework, despite its efficiency, requires an additional computation step on the client side to estimate the local loss, which is then used to inform the adaptive quantization level [9]. Furthermore, as highlighted in surveys on the topic, it is often impractical to measure every client's utility in every single round, as this information can frequently only be determined after a client has already participated in training, creating a difficult circular dependency [11]. The computational burden of executing these sophisticated scheduling algorithms on the server also cannot be ignored, raising the question of whether the efficiency gains truly outweigh the added system complexity.

3.1.3 Emerging Concerns in Fairness, Robustness, and Security

Finally, these new paradigms introduce new concerns regarding fairness, robustness, and security. Client selection algorithms that prioritize clients based on utility inherently favor those with high-quality data and superior system resources. This creates a significant risk of marginalizing "straggler" clients with limited resources or non-mainstream data, leading to a biased global model that neglects underrepresented populations and thus exacerbates fairness issues [11]. On the other hand, decentralized topologies, while eliminating the single point of failure of a central server, also forfeit its role as a trusted coordinator and security checkpoint. This architectural shift exposes the system to a greater risk of peer-to-peer security threats. As comprehensive surveys on decentralized learning point out, the absence of a central arbiter makes the system more vulnerable to a range of attacks, including data poisoning and privacy inversion, as malicious actors can interact directly with honest clients [14]. Even knowledge distillation is not immune, as attackers can inject malicious knowledge through poisoned logits, which are harder to detect than corrupted model parameters, or potentially reverse-engineer private data from the shared logit information [13].

3.2. Future Prospects

Based on these challenges, the path forward is to develop more integrated, intelligent, and trustworthy systems. Future research should synthesize existing methods into holistic frameworks that can dynamically adapt to heterogeneous environments while upholding fairness and security.

3.2.1 Hybrid, Cross-Layer Communication Optimization Frameworks

A significant opportunity lies in designing hybrid frameworks that synergistically combine strategies from all three categories. Since each category addresses the communication bottleneck from a different angle, a multi-layered optimization strategy could yield compounded efficiency gains. A promising direction is to develop a hierarchical system that operates at different levels of granularity. For instance, a framework could first use a decentralized or clustered topology to form collaborative groups of similar clients, a concept explored in hybrid DFL architectures [14]. Within each cluster, a resource-aware client selection mechanism could then be applied to select the most suitable participants for a given round. Finally, these selected clients could communicate using a highly efficient parameter-efficient learning technique, such as exchanging only the updates for lightweight LoRA modules. The principle of combining different adaptive strategies has already shown its effectiveness in a narrower context, such as combining time- and client-adaptive quantization [9], suggesting that a broader, cross-category integration is a logical and powerful next step.

3.2.2 Lightweight and Automated Adaptivity

To counteract the overhead of complex adaptive strategies, future work should focus on more lightweight and automated mechanisms. Instead of relying on costly device profiling, systems could be designed to implicitly infer client state from lightweight proxy metrics, like update variance or communication round-trip time. A more advanced direction is to automate the decision-making process itself. The challenge of balancing the selection of high-utility clients with the need to explore the wider client population is a classic exploitation-exploration dilemma, which strongly suggests the applicability of reinforcement learning (RL) [11]. Future frameworks could employ an RL agent on the server to learn an optimal, context-aware client selection policy over time, adapting to changes in the network and client population without requiring a manually engineered utility function. The success of methods that automate the tuning of specific hyperparameters, such as dynamically adjusting the number of local epochs to minimize communication rounds, demonstrates the feasibility and potential of this automated approach [10].

3.2.3 Next-Generation Paradigms Balancing Efficiency, Fairness, and Security

While paramount, communication efficiency cannot come at the expense of fairness and security. A critical research direction is the development of next-generation paradigms that explicitly co-design for these multiple objectives. To address the fairness concerns of utility-based client selection, new algorithms can be developed that incorporate long-term fairness metrics, ensuring underrepresented clients are not starved of participation. For decentralized learning, where the lack of a central authority creates security vulnerabilities, integrating technologies like blockchain could establish a trusted, server-less framework for identity management, contribution tracking, and incentive mechanisms, as proposed in the DFL literature [14]. Finally, to mitigate the privacy risks associated with knowledge distillation, future work can explore methods for fortifying the privacy guarantees of shared knowledge carriers. As suggested by recent surveys, techniques such as applying differential privacy to the logits before transmission could prevent the leakage of sensitive information, thus creating a more robust and privacy-preserving knowledge sharing protocol [13].

4. Conclusion

This survey has systematically addressed the critical challenge of communication efficiency in Personalized Federated Learning (PFL), a pivotal factor for its practical deployment in heterogeneous environments. By providing a structured taxonomy, this paper has illuminated a spectrum of strategies that tackle this challenge from distinct angles—by redesigning the model architecture for inherent efficiency, by optimizing the communication process with adaptive intelligence, and by fundamentally rethinking the collaborative network topology. The analysis highlights the inherent trade-offs between personalization fidelity and communication overhead, the system complexity of dynamic methods, and crucial emerging concerns regarding fairness, robustness, and security.

Looking ahead, the path forward lies in developing holistic, integrated frameworks. Future work must co-design for these multi-faceted objectives, creating intelligent systems that synergistically balance communication efficiency with model performance, equity, and security to unlock the true potential of distributed, personalized AI.

References

- [1] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag* 2020; 37 (3): 50–60.
- [2] McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: *Proc. Int. Conf. Artificial Intelligence and Statistics*. PMLR; 2017.
- [3] Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning. In: *Proc. 2020 4th World Conf. Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE; 2020.
- [4] Tan AZ, Yu H, Zhang L, Yang Q. Towards personalized federated learning. *IEEE Trans Neural Netw Learn Syst* 2022; 34 (12): 9587–9603.
- [5] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. *ACM Trans Intell Syst Technol* 2019; 10 (2): 1–19.
- [6] Ma X, Wang L, Zhang Z, Li J, Zhang Y. Layer-wised model aggregation for personalized federated learning. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. 2022.
- [7] Wang Z, Chen X, Zhu H, Zhang R, Zhang J. Federated low-rank adaptation for large language model fine-tuning over wireless networks. In: *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. IEEE; 2024.
- [8] Wang T, Chen J, Sun J, Li Y. Personalized federated learning via heterogeneous modular networks. In: *Proc. IEEE Int. Conf. Data Mining (ICDM)*. IEEE; 2022.
- [9] Hönig R, Zhao Y, Mullins R. DAdaQuant: Doubly-adaptive quantization for communication-efficient federated learning. In: *Proc. Int. Conf. Machine Learning (ICML)*. PMLR; 2022.
- [10] Xiang T, Xu X, Chen W, Li Y. Federated learning with dynamic epoch adjustment and collaborative training in mobile edge computing. *IEEE Trans Mobile Comput* 2023; 23 (5): 4092–4106.
- [11] Fu L, Zhang Z, Li H, Liu X. Client selection in federated learning: Principles, challenges, and opportunities. *IEEE Internet Things J* 2023; 10 (24): 21811–21819.
- [12] Yan Y, Tong X, Wang S. Clustered federated learning in heterogeneous environment. *IEEE Trans Neural Netw Learn Syst* 2023; 35 (9): 12796–12809.
- [13] Pang X, Li X, Liu H, Zhang T. When federated learning meets knowledge distillation. *IEEE Wireless Commun* 2024; 31 (5): 208–214.
- [14] Yuan L, Zhang P, Wang D, Xu C. Decentralized federated learning: A survey and perspective. *IEEE Internet Things J* 2024; 11 (21): 34617–34638.