

Energy-Efficient Federated Learning for Distributed Edge–Cloud Computing Environments

Suresh Geripathi

Department of Computer Science, University of North Texas, Denton, TX, USA.
contactsuresh@unt.edu

Krish Gaha

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
krishs@uc.edu

Abstract

Federated learning enables distributed model training across edge devices and cloud resources without transferring raw data to a central repository. However, the energy implications of periodic communication, heterogeneous hardware, unbalanced participation, and continuous model aggregation remain insufficiently addressed in production edge-cloud environments. This paper presents a system-level analysis of energy-efficient federated learning, emphasizing architectural trade-offs, resource governance, deployment policy, and sustainability. The central argument is that energy efficiency cannot be achieved solely through algorithmic compression or hardware acceleration; it must be designed as an emergent property of the interaction among protocol design, edge-cloud topology, operational incentives, and fairness constraints. The paper examines the structural differences between centralized and hierarchical federated architectures, explores communication-computation trade-offs, and discusses how serverless cloud orchestration and edge resource management affect energy consumption. It further addresses robustness and fairness, showing that energy optimization strategies may create uneven participation burdens on battery-powered devices and underrepresented nodes. Governance frameworks are considered through the lens of shared infrastructure management, carbon accounting, and multi-stakeholder coordination. The analysis connects federated learning to broader sustainable machine learning debates and offers forward-looking research directions in carbon-aware scheduling, energy reporting standards, and policy-aware system design. Rather than proposing a single optimization algorithm, the paper contributes a conceptual foundation for designing federated learning infrastructures that balance training quality, privacy, equity, and environmental responsibility.

Keywords

federated learning; edge computing; cloud computing; energy efficiency; distributed systems; sustainability; governance.

1. Introduction

Federated learning has emerged as a distributed training paradigm that allows model improvements to occur without centralizing raw data. This approach addresses privacy, communication cost, and data sovereignty concerns in edge-rich environments. The foundational formulation of federated averaging established that participating clients can perform local optimization and transmit only model updates to a coordinating server, thereby reducing raw data movement [1]. Subsequent surveys have positioned federated learning as a central component of privacy-preserving artificial intelligence, while also identifying

heterogeneous systems, energy budgets, and non-independent data distributions as persistent constraints [2]. At the same time, the deployment of learning workloads across distributed edge-cloud infrastructures introduces a complex socio-technical challenge in which energy consumption, latency, fairness, and governance interact. A comprehensive research agenda on federated learning has emphasized the need to move beyond accuracy-centric evaluation and toward system-level objectives that include communication efficiency, environmental footprint, and fairness [3].

Edge computing extends computational capability closer to data sources such as sensors, vehicles, and industrial controllers. This proximity reduces the need to transfer large volumes of raw telemetry to centralized facilities, but it also fragments computational capacity and introduces heterogeneous energy profiles. Early edge computing research articulated the structural tension between localized processing and central coordination [4]. More recently, serverless cloud orchestration for distributed machine learning has been examined as a way to manage dynamic training workloads through fine-grained resource allocation, with implications for energy consumption in cloud-based training pipelines [5]. The emergence of edge computing as a first-class infrastructure layer has further enabled low-latency services, but it requires careful design of resource scheduling and model synchronization [6]. Edge intelligence studies have shown that artificial intelligence workloads can be decomposed across cloud and edge tiers to balance accuracy, delay, and energy use [7].

The present paper examines energy-efficient federated learning from a systems perspective. It does not focus on a single optimization algorithm; rather, it analyzes architectural choices, communication and computation trade-offs, infrastructure governance, robustness, fairness, sustainability, deployment, and policy. The core position is that energy efficiency in federated learning is not merely a technical optimization target but a structural property that emerges from the interaction of protocols, hardware, network topology, organizational policy, and environmental constraints.

2. Background and Related Work

Federated learning has moved from theoretical privacy-preserving training to large-scale system implementations. In-edge artificial intelligence architectures have combined federated learning with mobile edge caching and communication to reduce cloud dependency [8]. Surveys of mobile edge computing emphasize the communication perspective and show that wireless uplink and downlink conditions dominate end-to-end learning latency and energy consumption [9]. Federated machine learning has also been formalized as a general distributed learning concept with a focus on data owners retaining control, but this literature simultaneously warns that naive federation may transfer significant communication overhead [10]. At the systems level, federated learning at scale requires device selection, secure aggregation, and asynchronous participation because real devices are unreliable and constrained [11].

Energy awareness has gained attention in machine learning more broadly. In natural language processing, training large models entails substantial electricity use and carbon emissions, prompting calls for efficiency reporting and policy-informed model development [12]. Measurement studies have quantified carbon emissions of machine learning workloads and demonstrated that location, hardware, and training time interact to shape environmental impact [13]. In climate-oriented machine learning research, computational efficiency is presented as one dimension of responsible system design, alongside applications in energy forecasting, transportation, and land use [14]. These broader findings apply directly to

federated learning, where many edge devices operate under battery or thermal constraints and where cloud aggregation nodes may run continuously.

This paper departs from algorithm-centric surveys by treating energy efficiency as a socio-technical systems property. While existing work has addressed compression, sparsification, and quantization, the structural implications of those choices for infrastructure governance and fairness remain underexplored.

3. Architectural Considerations for Energy-Aware Federated Learning

Energy-efficient federated learning depends first on how the distributed system is decomposed across cloud and edge layers. A fully centralized architecture places aggregation in a cloud data center, with edge clients performing local updates. This design simplifies coordination but can create high wide-area network energy costs when model parameters are large or when many clients participate in each round. A hierarchical architecture introduces intermediate edge aggregators, which can reduce the distance between clients and coordination points, consolidate updates, and lower backbone traffic. A comprehensive survey of federated learning in mobile edge networks has shown that these architectural choices affect not only communication energy but also the ability to satisfy latency constraints under heterogeneous connectivity [15]. Similarly, deep learning with edge computing has been characterized as a continuum in which computation can shift between cloud, fog, and device layers depending on workload demands and energy availability [16].

The placement of aggregation functions has direct energy implications. Cloud-based aggregation benefits from high availability and strong computational capacity, but it may force resource-limited clients to transmit updates over long distances. Edge-based aggregation reduces long-haul communication and can exploit local energy sources, yet it introduces additional coordination complexity and may require more capable edge nodes. In practice, many deployments adopt a hybrid model in which local aggregators serve geographic or organizational clusters and a cloud coordinator handles global model versioning. This hybrid structure allows operators to align energy consumption with local conditions, such as renewable energy availability, battery state, and network congestion.

Infrastructure heterogeneity is another structural factor. Edge devices range from smartphones and wearable sensors to embedded industrial controllers, each with distinct processor capabilities, memory limits, thermal envelopes, and power supplies. Cloud data centers exhibit more homogeneous hardware but differ in power usage effectiveness, cooling efficiency, and carbon intensity. A system-level energy analysis must therefore account for the distribution of computation across these tiers, because shifting work from one tier to another can reduce communication energy while increasing computation energy or vice versa. Architectural decisions should be evaluated not solely by model accuracy, but by the combined energy profile of local computation, intermediate aggregation, global training, and data transfer.

Serverless cloud orchestration adds a further architectural dimension. Function-based scheduling can allocate aggregation and preprocessing tasks dynamically, reducing idle capacity and improving resource utilization [5]. However, serverless execution may also introduce cold-start overhead and limited state persistence, which can interact with federated aggregation in complex ways. An energy-aware architecture must therefore consider whether aggregation functions run as continuously provisioned services or as transient serverless tasks, and how that choice affects both energy consumption and operational reliability.

4. Communication and Computation Trade-offs

The energy cost of federated learning is shaped by a continuous negotiation between communication and local computation. Each global round requires clients to receive the current model, perform local training, and transmit updates. If clients perform many local epochs between communication rounds, the total communication frequency decreases, but local computation energy increases and the convergence behavior may change. Conversely, frequent communication can improve synchronization but raises wireless transceiver energy and backbone network load. This trade-off is especially pronounced in mobile edge environments where wireless uplink conditions vary significantly across clients and over time [9]. System designers therefore need to select participation schedules, update compression methods, and local training budgets that match the energy constraints of the most constrained nodes, not only the average node.

Model compression and sparsification techniques are commonly proposed to reduce communication energy. By transmitting only important gradients or quantized update vectors, federated systems can lower the number of bytes sent over wireless and wide-area links. However, these techniques can introduce approximation errors that may slow convergence or degrade fairness across heterogeneous data distributions. The energy saved during communication may be partially offset by additional local computation required to generate compressed updates or by more training rounds needed to reach the same model quality. From a system perspective, compression should be treated as an energy redistribution mechanism rather than an unambiguous efficiency improvement.

Asynchronous federated protocols present another trade-off. Synchronous aggregation requires waiting for the slowest selected participants, which can increase idle energy on faster devices and extend wall-clock training time. Asynchronous methods allow faster devices to proceed, but they can introduce stale updates and complicate convergence analysis. For energy efficiency, asynchronous participation may reduce idle waiting and permit devices to train when they have available energy, such as when a smartphone is charging or when an industrial sensor enters a low-load period. At the same time, asynchronous systems require more sophisticated versioning and consistency mechanisms, which themselves consume computational and organizational resources.

The communication-computation trade-off also intersects with data heterogeneity. When client data distributions differ, local updates may diverge, and additional communication may be needed to prevent model degradation. Compression that is too aggressive may amplify this divergence and produce unequal model performance across participating sites. Therefore, energy-efficient protocol design must integrate compression, scheduling, and heterogeneity management rather than optimize communication cost in isolation. A system-level approach evaluates the combined energy of local processing, wireless transmission, aggregation, and model distribution across the full training lifecycle.

5. Robustness, Fairness, and Equity

Energy optimization strategies in federated learning can have distributional consequences. Devices with large batteries or reliable power sources may be selected more frequently, while battery-constrained devices may be excluded or penalized if the optimization objective emphasizes aggregate energy reduction. This can create fairness concerns, particularly when the excluded devices belong to underrepresented populations or operate in environments with less reliable energy infrastructure. Fairness in machine learning has been defined through

statistical criteria that compare prediction outcomes across groups [17]. In federated settings, these concerns extend beyond prediction accuracy to include participation burdens, data contribution incentives, and access to model benefits. Individual fairness formulations further suggest that similar participants should experience similar treatment, which is difficult to guarantee when device capabilities and energy costs differ widely [18]. A broader fairness framework for machine learning systems emphasizes that technical objectives cannot be separated from the social contexts in which systems are deployed [19].

Robustness is also closely linked to energy management. Federated systems must tolerate stragglers, intermittent connectivity, and device churn. If an energy-saving policy reduces the number of participating clients too aggressively, the global model may become less robust to rare but important data patterns. Conversely, including too many low-energy devices may cause frequent dropouts and destabilize training. Robustness therefore requires designing participation and aggregation mechanisms that maintain model quality while respecting heterogeneous energy budgets. This may involve allowing delayed contributions, weighting updates according to reliability, or maintaining multiple model versions for different device groups.

Fairness and energy efficiency can sometimes be aligned through incentive design. Devices that contribute more computational work could receive credits, monetary compensation, or improved model access. However, incentive mechanisms may also encourage participants to overstate their energy capacity or to manipulate update contributions. Transparent accounting of energy costs and contributions can help build trust across organizations and reduce opportunistic behavior. Governance structures that include participants in decision-making about scheduling, compression, and participation thresholds are better equipped to balance energy efficiency with fairness than purely centralized optimization.

6. Edge-Cloud Infrastructure and Resource Governance

Energy-efficient federated learning operates within shared infrastructure that spans organizational boundaries. Cloud providers, network operators, edge facility owners, and client device administrators each influence resource allocation. This multi-stakeholder setting requires governance arrangements that define responsibilities, energy budgets, access rules, and accountability. Institutional analysis of shared resources suggests that sustainable management depends on clear boundaries, collective choice arrangements, monitoring, and graduated sanctions [20]. Applied to federated learning, such principles imply that energy efficiency cannot be achieved by a single operator acting alone, but requires coordinated rules across infrastructure layers.

Carbon-aware scheduling is one governance mechanism that can align federated training with environmental objectives. Cloud and edge resources located in regions with high renewable energy availability can be prioritized for aggregation and large-batch processing. Conversely, tasks that are not time-sensitive can be delayed until carbon intensity is lower. However, carbon-aware scheduling may conflict with latency requirements or with fairness goals if it systematically favors certain geographic regions. Governance frameworks must therefore specify how environmental objectives are weighted against service quality and equity.

Energy accounting is another governance challenge. Federated learning workloads are distributed across many devices, making it difficult to measure total energy consumption accurately. Cloud providers may report data center energy use, but client-side energy consumption is often estimated rather than directly measured. Without standardized

accounting, organizations cannot reliably compare different architectural choices or assess the environmental benefit of compression and scheduling techniques. Energy reporting standards for federated systems would improve transparency and enable better procurement and policy decisions.

Multi-tenancy further complicates resource governance. Edge nodes may serve multiple applications simultaneously, including video analytics, network functions, and federated learning tasks. Resource contention can increase energy consumption and degrade performance. Governance arrangements should include resource isolation policies, priority classes, and conflict resolution procedures. In some cases, federated learning can be treated as a background workload that uses spare capacity, but this requires careful management to avoid unintended interference with safety-critical services.

7. Deployment Models and Operational Policy Implications

Deployment models for federated learning vary widely across sectors. In smart manufacturing, edge controllers may participate in local model training while a cloud platform coordinates cross-facility learning. In healthcare, hospitals may operate their own edge aggregators to preserve patient privacy while contributing to a shared research model. In smart cities, municipal infrastructure may host intermediate aggregators that consolidate updates from traffic sensors, environmental monitors, and public safety systems. Each deployment model has different energy implications depending on the density of devices, the availability of wired backhaul, and the criticality of the application.

Operational policy should specify when training occurs and how devices are selected. For example, federated learning on smartphones can be scheduled during overnight charging periods to reduce battery stress and take advantage of lower carbon intensity in some regions. Industrial deployments may schedule training during maintenance windows or periods of low production activity. These policies require coordination between organizational units that manage devices, networks, and cloud resources. A purely technical scheduling algorithm cannot enforce such policies without institutional support.

Service level agreements for federated learning should include energy-related metrics, not only accuracy and latency. Organizations may agree on maximum communication budgets, minimum participation fairness, or carbon intensity thresholds. Such agreements can drive architectural choices and encourage investment in energy-efficient hardware and networking equipment. Policy instruments such as carbon pricing, energy efficiency standards, and public procurement criteria can further incentivize the adoption of energy-aware federated systems. However, these instruments must be designed carefully to avoid imposing disproportionate costs on smaller organizations or on edge environments with limited access to renewable energy.

The operational lifecycle of federated learning also includes model evaluation, drift monitoring, and retraining. Each of these stages consumes energy and should be included in system-level energy assessments. Continuous monitoring may require frequent inference on edge devices, which can be as energy-intensive as training when deployed at scale. A comprehensive deployment policy therefore addresses not only training rounds but also inference, data preprocessing, secure aggregation, and model distribution.

8. Future Research Directions

Future research should develop integrated energy models that capture computation, communication, cooling, and device-specific overheads across the full federated learning lifecycle. These models should account for heterogeneous hardware, dynamic network conditions, and renewable energy availability. Rather than optimizing a single metric, researchers should explore multi-objective evaluation frameworks that include accuracy, energy, latency, fairness, and robustness. Such frameworks would support comparative analysis across compression methods, scheduling algorithms, and architectural configurations.

Benchmarking is another important direction. Public datasets and reference workloads for federated learning exist, but they rarely include detailed energy measurements for edge devices and cloud aggregators. Energy-aware benchmark suites would enable reproducible comparisons and accelerate the development of efficient protocols. These benchmarks should include realistic device heterogeneity, intermittent connectivity, and non-independent data distributions.

Governance research should examine how institutional rules shape energy outcomes in federated systems. Case studies of cross-organizational federated deployments can reveal how decisions about data sharing, resource access, and cost allocation influence energy consumption. Comparative analysis across sectors such as healthcare, manufacturing, and public infrastructure would help identify governance patterns that support both efficiency and equity. Researchers should also investigate how carbon accounting standards can be adapted to distributed learning workloads that span multiple jurisdictions.

Finally, the relationship between energy efficiency and model utility requires deeper study. Aggressive energy reduction may reduce model quality for some participants, potentially undermining the value of federation. Research is needed on fairness-aware energy allocation, in which energy budgets are distributed according to participant needs and contributions rather than solely according to aggregate efficiency. This direction aligns with broader calls for sustainable and socially responsible machine learning systems [14]. It also connects federated learning to infrastructure governance scholarship and to emerging policy debates about the environmental footprint of artificial intelligence.

9. Conclusion

Energy-efficient federated learning in distributed edge-cloud environments is a systems challenge that spans algorithms, architectures, infrastructure governance, and policy. Communication and computation trade-offs can be shaped through compression, hierarchical aggregation, and asynchronous scheduling, but each choice redistributes energy burdens across participants and infrastructure layers. Robustness and fairness concerns are inseparable from energy optimization because energy-saving policies can exclude resource-constrained devices and create uneven participation patterns. Governance arrangements across cloud providers, edge operators, and client organizations are therefore essential to align technical choices with environmental and equity objectives. The adoption of carbon-aware scheduling, transparent energy accounting, and energy-sensitive service level agreements can improve both sustainability and operational accountability. Future work should integrate energy measurement into federated learning benchmarks, develop multi-objective evaluation frameworks, and examine how institutional rules shape energy outcomes in real deployments. By treating energy efficiency as an emergent property of distributed socio-technical systems rather than as a narrow optimization target, researchers and practitioners can build federated learning infrastructures that are not only accurate and private but also environmentally responsible and socially equitable.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.
2. 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.
3. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., . . . Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
4. 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.
5. Chen, X. (2025, July). Research on Cloud Computing Applications for Optimizing DML Model Training in Serverless Architecture. In *International Conference on Frontier Computing* (pp. 553-562). Singapore: Springer Nature Singapore.
6. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30–39.
7. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8), 1738–1762.
8. Wang, X., Han, Y., Wang, C., Zhao, Q., Chen, X., & Chen, M. (2019). In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network*, 33(5), 156–165.
9. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys & Tutorials*, 19(4), 2322–2358.
10. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1–19.
11. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1, 374–388.
12. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
13. Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.
14. Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D.,

- Platt, J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2022). Tackling climate change with machine learning. *ACM Computing Surveys*, 55(2), 1–96.
15. Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., Niyato, D., & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3), 2031–2063.
 16. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8), 1655–1674.
 17. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323.
 18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
 19. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.
 20. Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, 325(5939), 419–422.