

Energy-Aware Scheduling of Collaborative AI Agents for Efficient Edge Computing and Internet of Things Applications

Alberico Baroffio

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
alberico468@ucf.edu

Santeri Nikula

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
contactsanteri@unh.edu

Abstract

The convergence of edge computing, the Internet of Things, and collaborative artificial intelligence creates new possibilities for low-latency, context-sensitive services, but it also introduces substantial energy and coordination challenges. This paper examines energy-aware scheduling of collaborative AI agents as a system-level problem that spans infrastructure, middleware, governance, and sustainability. Rather than treating scheduling as a narrow optimization task, the analysis positions it within the broader architecture of edge intelligence, where heterogeneous devices, intermittent connectivity, privacy constraints, and multi-tenant workloads must be reconciled simultaneously. The paper argues that collaborative AI agents can improve resource utilization by decomposing tasks, sharing context, and adapting to changing conditions, yet these benefits depend on scheduling policies that account for energy budgets, fairness, robustness, and policy compliance. Architectural patterns such as hierarchical orchestration, federated coordination, and hybrid cloud-edge-fog deployment are compared in terms of their structural trade-offs. The discussion also considers governance mechanisms, including transparency, accountability, and equitable access, as necessary conditions for responsible deployment. Case illustrations from smart manufacturing, intelligent transportation, environmental monitoring, and public safety demonstrate how energy-aware scheduling affects service quality, operational cost, and environmental impact. The paper concludes that efficient edge and IoT ecosystems require adaptive scheduling frameworks that combine technical optimization with institutional and policy design. Future research should address cross-layer energy accounting, resilient coordination under failure, and fair resource allocation across diverse stakeholders.

Keywords

edge computing, Internet of Things, collaborative AI agents, energy-aware scheduling, sustainability, governance, distributed systems, fairness, robustness.

1. Introduction

The rapid expansion of edge computing and the Internet of Things has shifted computation closer to data sources, enabling services that demand low latency, high bandwidth efficiency, and contextual awareness [1]. In parallel, artificial intelligence has moved from centralized cloud environments toward distributed edge nodes, where models must operate under constrained energy, memory, and connectivity conditions [2]. This movement has created a

new class of socio-technical infrastructure in which autonomous and semi-autonomous agents collaborate across devices, gateways, and micro-data centers. The resulting systems are not merely technical assemblies; they are embedded in urban environments, industrial processes, transportation networks, and public services. Consequently, scheduling decisions have consequences that extend beyond throughput and response time to include energy consumption, environmental sustainability, fairness, and governance.

The Internet of Things provides the sensing and actuation substrate for these systems, connecting heterogeneous devices that range from simple sensors to complex machines [9]. Edge computing supplies proximate computation, storage, and networking resources, while cloud platforms provide elastic capacity for training, long-term analytics, and coordination [10]. Collaborative AI agents introduce an additional layer of complexity because they can negotiate, delegate, and coordinate tasks without direct human intervention. Energy-aware scheduling of such agents is therefore a multidimensional problem. It must decide which agent should execute which task, when migration is beneficial, how much communication is acceptable, and how to balance local autonomy with global objectives. These decisions are shaped by physical constraints, economic incentives, regulatory requirements, and social expectations.

This paper develops a system-level analysis of energy-aware scheduling for collaborative AI agents in edge and IoT environments. The central argument is that energy efficiency cannot be achieved through isolated algorithmic improvements alone. Instead, it requires architectural alignment, governance mechanisms, robust deployment strategies, and policy frameworks that recognize the interdependencies among computation, communication, and human institutions. The paper is organized as follows. Section two reviews the foundations of edge intelligence and collaborative agents. Section three frames energy-aware scheduling as a system-level problem. Section four examines architectural patterns and their trade-offs. Section five discusses governance, fairness, and policy implications. Section six analyzes deployment, robustness, and sustainability. Section seven provides cross-domain comparisons and case illustrations. Section eight identifies open challenges and research directions. Section nine concludes with implications for future research and practice.

2. Foundations of Edge Intelligence and Collaborative Agents

Edge intelligence refers to the integration of artificial intelligence capabilities with edge computing infrastructure so that inference, learning, and decision-making can occur near data sources [5]. This integration is motivated by latency requirements, privacy concerns, bandwidth limitations, and the need for continuous operation during network disruptions. Early edge intelligence systems often relied on static model deployment, but contemporary approaches increasingly emphasize adaptive model placement, distributed inference, and collaborative learning [6]. The movement toward collaboration reflects the recognition that no single device possesses sufficient energy, computation, or data to address complex tasks independently. Instead, multiple agents can share intermediate results, divide workloads, and collectively adapt to environmental changes.

The concept of an agent has a long history in artificial intelligence and multi-agent systems. Intelligent agents are typically characterized by autonomy, reactivity, proactivity, and social ability [11]. In distributed systems, agents can be designed to pursue local goals while participating in broader coordination protocols [12]. Multi-agent systems research has explored negotiation, coalition formation, task allocation, and learning in dynamic environments [13]. More recent work has extended these ideas to networked and resource-

constrained settings, including IoT and edge domains [14]. Collaborative AI agents in edge environments inherit these properties but operate under stricter energy and communication constraints. They must often make decisions with incomplete information, intermittent connectivity, and heterogeneous hardware.

The distinction between centralized and distributed intelligence is important for scheduling. Centralized orchestration can achieve global consistency and simplify policy enforcement, but it may create bottlenecks and single points of failure. Distributed coordination can improve resilience and local responsiveness, but it complicates energy accounting and fairness guarantees. Hybrid approaches attempt to combine the strengths of both by assigning strategic decisions to regional controllers and tactical decisions to local agents. The effectiveness of such approaches depends on the granularity of state information, the frequency of coordination, and the cost of communication. In energy-aware scheduling, these factors are not secondary; they determine whether collaboration yields net benefits or simply shifts energy consumption from computation to communication.

3. Energy-Aware Scheduling as a System-Level Problem

Energy-aware scheduling in edge and IoT systems involves allocating tasks, data, and communication resources in ways that satisfy service requirements while minimizing energy consumption or respecting energy budgets. In cloud data centers, energy-aware resource allocation has been studied extensively through consolidation, dynamic voltage scaling, and workload migration [17]. Edge environments differ because resources are geographically distributed, often powered by batteries or constrained energy sources, and subject to variable connectivity [16]. Mobile edge computing introduces additional dynamics related to user mobility, radio access conditions, and backhaul availability [3]. These factors mean that scheduling policies must be adaptive, context-aware, and capable of operating under uncertainty.

A central trade-off is the relationship between computation and communication energy. Moving a task to a more powerful edge server may reduce execution time but increase data transmission energy. Conversely, executing locally may conserve communication energy but drain a battery-constrained device. The optimal choice depends on task characteristics, model size, network conditions, and the energy profiles of the participating nodes [4]. Collaborative AI agents can improve this trade-off by decomposing tasks into subtasks that can be processed at different locations. For example, a vision agent might perform initial filtering locally, send compressed features to a nearby gateway, and rely on a regional server for complex inference. Such decomposition requires scheduling mechanisms that understand dependencies, deadlines, and the energy cost of each stage.

Deep learning workloads add further complexity because inference and training have different energy profiles [8]. Inference is often periodic and latency-sensitive, while training is computationally intensive and can be delayed or batched. Collaborative agents may engage in federated learning, where local updates are computed on devices and aggregated at edge or cloud nodes. Scheduling in this context must balance model accuracy, communication overhead, privacy, and energy consumption. If too many devices participate, communication energy may dominate. If too few participate, model quality may degrade. Energy-aware scheduling therefore becomes a multi-objective problem that cannot be reduced to a single metric. It must consider service-level agreements, fairness among users, robustness to failures, and long-term sustainability.

4. Architectural Patterns for Collaborative Agent Scheduling

Several architectural patterns have emerged for coordinating collaborative AI agents in edge and IoT systems. Hierarchical orchestration uses a layered control structure in which cloud, fog, and edge nodes have different responsibilities. Strategic planning and global policy enforcement occur at higher layers, while real-time scheduling and local adaptation occur at lower layers. This pattern aligns with autonomic computing principles, where systems manage themselves according to high-level objectives [15]. Hierarchical orchestration simplifies governance and can improve scalability, but it may introduce latency in decision-making and reduce local autonomy. It also requires reliable upward and downward communication, which may not be available in remote or disrupted environments.

Federated coordination distributes decision-making across peer agents that negotiate and share information without a central controller. This pattern can improve resilience and respect local constraints, but it complicates global energy optimization and fairness. Agents may pursue local objectives that conflict with system-wide sustainability goals. Recent scaling work on agentic models for full-scenario applications suggests that larger and more general agent architectures can support broader task coverage, yet such models may increase energy demand if deployed without careful scheduling [18]. The challenge is to design coordination protocols that allow agents to specialize, share context, and avoid redundant computation. Fog computing architectures provide an intermediate layer that can host such coordination functions closer to the edge than cloud data centers [19]. The placement of coordination functions affects both energy consumption and response time.

Hybrid patterns combine hierarchical and federated elements. For example, a regional controller may set energy budgets and policy constraints, while local agents negotiate task assignments within those boundaries. This approach can balance global oversight with local adaptability. However, it requires robust mechanisms for monitoring, auditing, and conflict resolution. In multi-tenant environments, different applications may have different priorities and energy profiles. A smart factory may prioritize safety-critical tasks, while a consumer application may prioritize cost. Scheduling policies must therefore be expressive enough to represent diverse objectives and transparent enough to justify allocation decisions. Architectural choices are not merely technical; they determine who has control over resources and how accountability is distributed.

5. Governance, Fairness, and Policy Implications

Governance of energy-aware scheduling involves decisions about access, prioritization, transparency, and accountability. In edge and IoT ecosystems, resources are often shared among public institutions, private firms, and individual users. If scheduling policies systematically favor powerful actors, they can reinforce existing inequalities. Fairness may be understood in terms of equal access, proportional allocation, or need-based prioritization. Each interpretation leads to different scheduling outcomes. For example, a hospital network may require guaranteed low-latency capacity for emergency applications, while residential users may accept periodic delays in exchange for lower energy consumption. Policy frameworks must therefore specify the principles that guide trade-offs between efficiency and equity.

Security and privacy are closely linked to governance. Distributed IoT systems are vulnerable to attacks that can manipulate scheduling decisions, falsify energy data, or disrupt coordination [20]. If agents rely on untrusted information, they may make inefficient or

unsafe choices. Blockchain and distributed ledger technologies have been proposed to strengthen trust, traceability, and transactional integrity in IoT environments [21]. While these mechanisms can support auditability, they also introduce computational and energy overhead. Governance must therefore consider whether the benefits of trust mechanisms outweigh their costs. In energy-constrained settings, lightweight attestation and reputation systems may be more appropriate than computationally intensive consensus protocols.

Policy implications extend beyond technical standards. Governments and regulators may need to establish energy reporting requirements for edge infrastructures, similar to efficiency standards for data centers. They may also need to address cross-border data flows, because scheduling decisions can move workloads across jurisdictions with different privacy and energy regulations. Collaborative AI agents complicate accountability because decisions emerge from interactions among multiple actors. If an agent causes harm or consumes excessive energy, it may be difficult to assign responsibility. Governance mechanisms should therefore include logging, explainability, and dispute resolution processes. Fairness and sustainability are not add-ons to scheduling; they are constitutive elements of legitimate and resilient socio-technical infrastructure.

6. Deployment, Robustness, and Sustainability

Deployment of energy-aware scheduling systems must account for heterogeneity, scale, and operational uncertainty. Edge devices vary widely in computation, memory, energy source, and connectivity [2]. Some are battery-powered sensors with limited processing, while others are gateways with stable power and substantial storage. Scheduling policies that assume homogeneous capabilities will fail in practice. They must adapt to device profiles, network conditions, and application requirements. Middleware plays a critical role by abstracting hardware differences and providing common interfaces for monitoring, task placement, and energy accounting [3]. Without such abstractions, collaborative agents cannot reliably coordinate across diverse environments.

Robustness is essential because edge and IoT systems operate in unpredictable conditions. Devices fail, networks partition, and energy sources fluctuate. A scheduling policy that optimizes average performance may perform poorly under stress. Robust scheduling should include redundancy, graceful degradation, and fallback strategies. For example, if a regional coordinator becomes unavailable, local agents should be able to continue operating with reduced coordination. If energy reserves fall below a threshold, agents should prioritize essential tasks and defer non-critical workloads. These behaviors require explicit policy design and testing. They also require energy models that reflect real-world variability rather than idealized assumptions.

Sustainability considerations extend beyond operational energy consumption. The manufacturing, deployment, and disposal of edge devices have environmental impacts. Extending device lifetimes through adaptive scheduling can reduce electronic waste. However, energy-aware scheduling may also encourage premature replacement if older devices cannot meet new model requirements. A life-cycle perspective is therefore necessary. Scheduling policies should consider not only instantaneous energy use but also embodied energy and material impacts. In large-scale IoT deployments, even small efficiency improvements can produce significant aggregate effects [5]. Conversely, inefficient coordination protocols can generate unnecessary communication and computation, undermining sustainability goals. Deployment strategies should therefore align technical optimization with broader environmental and social objectives.

7. Cross-Domain Comparisons and Case Illustrations

Smart manufacturing illustrates the tension between productivity and energy efficiency. In a factory, collaborative robots, sensors, and edge controllers must coordinate to maintain throughput while minimizing energy use. Scheduling decisions affect machine utilization, idle time, and peak demand. If agents can shift non-urgent tasks to periods of low energy prices or renewable availability, they can reduce costs and emissions. However, safety-critical tasks cannot be delayed. Governance mechanisms must define which tasks are flexible and which are not. Similar issues arise in intelligent transportation, where vehicles, roadside units, and traffic management centers collaborate. Energy-aware scheduling can reduce congestion and emissions, but it must also ensure fairness across neighborhoods and modes of transport [1].

Environmental monitoring provides a different perspective. Wireless sensor networks often operate in remote areas with limited energy harvesting. Collaborative agents can aggregate data, compress readings, and adapt sampling rates based on events. Scheduling must balance data quality with battery life. If too much data is transmitted, the network may fail prematurely. If too little is transmitted, important events may be missed. Edge intelligence enables local filtering and anomaly detection, reducing communication energy [7]. Fog computing can provide intermediate processing for regional analysis [19]. These architectures demonstrate how energy-aware scheduling can support long-term sustainability in resource-constrained settings.

Public safety and emergency response introduce urgent and unpredictable workloads. During disasters, communication infrastructure may be damaged, and energy supplies may be unreliable. Collaborative AI agents can coordinate search and rescue, monitor hazards, and allocate resources. Scheduling must prioritize life-critical tasks, maintain situational awareness, and adapt to rapidly changing conditions. Fairness concerns are acute because affected populations may have different levels of access to digital services. Governance must ensure that emergency response systems do not disproportionately neglect marginalized communities. These cases show that energy-aware scheduling is not only a technical optimization but also a matter of social and political judgment.

8. Open Challenges and Research Directions

Several open challenges remain. First, cross-layer energy accounting is difficult because energy consumption occurs at multiple levels: sensing, computation, communication, and actuation. Current models often focus on one layer and ignore interactions. A more integrated approach is needed to attribute energy costs accurately and to design incentives that align local and global objectives [6]. Second, resilience under adversarial conditions requires scheduling policies that can detect and respond to manipulation. If an attacker can falsify energy data or delay messages, coordination may fail. Robustness must be designed into both algorithms and architectures [8]. Third, scalability remains a concern as the number of agents and devices grows. Hierarchical and federated approaches must be evaluated at realistic scales.

Fourth, fairness in multi-tenant edge environments requires further research. Existing scheduling algorithms often optimize aggregate utility without considering distributional effects. Fairness metrics must be defined in ways that reflect diverse stakeholder values. Fifth, the energy implications of increasingly general agentic models must be examined. Recent work on scaling agentic models suggests that broader capabilities may come with higher computational demands [18]. Deployment strategies must therefore consider whether centralized or distributed inference is more sustainable for specific applications. Sixth,

governance mechanisms must evolve alongside technical systems. Blockchain and other trust technologies offer promising tools, but their energy costs and governance implications require careful assessment [21]. Multi-agent systems research can inform the design of coordination protocols that are both efficient and accountable [14].

Finally, interdisciplinary collaboration is essential. Computer scientists, engineers, policymakers, ethicists, and social scientists must work together to design edge and IoT systems that are efficient, fair, and sustainable. Technical metrics such as latency and energy consumption are necessary but insufficient. They must be complemented by indicators of equity, transparency, and environmental impact. The challenge is not merely to build faster or more efficient systems but to build systems that serve human needs without exacerbating existing inequalities or degrading the natural environment.

9. Conclusion

Energy-aware scheduling of collaborative AI agents for edge computing and IoT applications is a complex system-level challenge that cannot be addressed through algorithmic optimization alone. It requires architectural alignment, governance mechanisms, robust deployment strategies, and policy frameworks that recognize the interdependencies among computation, communication, energy, and society. Edge computing and IoT provide the infrastructure for distributed intelligence, while collaborative agents offer the flexibility needed to adapt to heterogeneous and dynamic conditions [1]. However, without careful scheduling, collaboration can increase energy consumption and create new forms of inequity. Recent advances in edge intelligence and agentic models expand the design space but also raise the stakes for responsible deployment [5] [18].

The analysis in this paper suggests that effective energy-aware scheduling must be multi-objective, context-aware, and institutionally embedded. It must balance efficiency with fairness, local autonomy with global coordination, and short-term performance with long-term sustainability. Hierarchical, federated, and hybrid architectures each have distinct trade-offs. Governance mechanisms must ensure transparency, accountability, and equitable access. Deployment strategies must account for heterogeneity, failure, and life-cycle impacts. Cross-domain comparisons demonstrate that these issues manifest differently across manufacturing, transportation, environmental monitoring, and public safety, yet they share common structural challenges. Future research should develop integrated energy models, resilient coordination protocols, fair allocation mechanisms, and policy frameworks that support sustainable and inclusive edge ecosystems.

References

1. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30-39. <https://doi.org/10.1109/MC.2017.9>
2. 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>
3. 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. <https://doi.org/10.1109/COMST.2017.2745201>

4. Abbas, N., Zhang, Y., Taherkordi, A., & Skeie, T. (2018). Mobile edge computing: A survey. *IEEE Internet of Things Journal*, 5(1), 450-465.
<https://doi.org/10.1109/JIOT.2017.2750180>
5. 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. <https://doi.org/10.1109/JPROC.2019.2918951>
6. Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(2), 869-904. <https://doi.org/10.1109/COMST.2020.2970550>
7. Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal*, 7(8), 7457-7469. <https://doi.org/10.1109/JIOT.2020.2984887>
8. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8), 1655-1674. <https://doi.org/10.1109/JPROC.2019.2921977>
9. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645-1660. <https://doi.org/10.1016/j.future.2013.01.010>
10. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4), 2347-2376.
<https://doi.org/10.1109/COMST.2015.2444095>
11. Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1(1), 7-38.
<https://doi.org/10.1023/A:1010090405266>
12. Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2), 115-152.
<https://doi.org/10.1017/S0269888900008122>
13. Stone, P., & Veloso, M. (2000). Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3), 345-383.
<https://doi.org/10.1023/A:1008178430446>
14. Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. *IEEE Access*, 6, 28573-28593. <https://doi.org/10.1109/ACCESS.2018.2831228>
15. Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, 36(1), 41-50. <https://doi.org/10.1109/MC.2003.1160055>
16. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), 599-616.
<https://doi.org/10.1016/j.future.2008.12.001>
17. Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems*, 28(5), 755-768.
<https://doi.org/10.1016/j.future.2011.04.017>

18. Jiang, Hao, et al. "AgentOmnia: Scaling Agentic Models for Full-Scenario Applications." arXiv preprint arXiv:2607.23124 (2026).
19. Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the Internet of Things. In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13-16). ACM. <https://doi.org/10.1145/2342509.2342513>
20. Roman, R., Zhou, J., & Lopez, J. (2013). On the features and challenges of security and privacy in distributed Internet of Things. *Computer Networks*, 57(10), 2266-2279. <https://doi.org/10.1016/j.comnet.2012.12.018>
21. Kshetri, N. (2017). Can blockchain strengthen the Internet of Things? *IT Professional*, 19(4), 68-72. <https://doi.org/10.1109/MITP.2017.3051335>