Asia Pacific Academy of Science Pte. Ltd.

Computer and Telecommunication Engineering 2025, 3(2), 8431.
https://doi.org/10.54517/cte8431

Article

Energy-efficient edge computing architecture for 5G networks: Evidence from real base station operational data

Zhongming Wang

School of Information Engineering, Changzhou University, Changzhou 213164, China; wangzm857@163.com

CITATION

Wang Z. Energy-efficient edge computing architecture for 5G networks: Evidence from real base station operational data. Computer and Telecommunication Engineering. 2025; 3(2): 8431. https://doi.org/10.54517/cte8431

ARTICLE INFO

Received: 7 March 2025
Accepted: 31 May 2025
Available online: 20 June 2025

Abstract: The rapid densification of fifth generation radio access networks and the growing demand for low-latency services have significantly increased the energy consumption of mobile infrastructures, raising critical concerns regarding operational cost and environmental sustainability. Multi-access edge computing has been introduced as a key architectural paradigm to support stringent latency requirements by deploying computing resources closer to base stations. However, the deployment of edge computing does not inherently guarantee energy efficiency, as edge platforms may consume substantial baseline power under low utilization if orchestration and task placement are not energy-aware. This paper proposes an energy-efficient edge computing architecture for 5G networks that integrates real-time energy monitoring with load-aware task scheduling at the edge layer. The proposed architecture is aligned with standardized 5G edge deployment frameworks and is evaluated using real operational base station data, including traffic load, computing utilization, and power consumption measurements. By leveraging real data rather than synthetic workloads, the proposed approach enables a realistic assessment of energy efficiency under practical operating conditions. Experimental results demonstrate that the proposed architecture achieves approximately 30% improvement in energy efficiency compared with a conventional edge computing deployment without energy-aware scheduling, while maintaining comparable latency performance. The findings indicate that data-driven energy-aware orchestration at the network edge can deliver measurable energy savings in commercial 5G environments. This work provides practical insights for mobile network operators seeking to reduce the energy footprint of 5G infrastructures and contributes a deployable architectural framework for energy-efficient edge computing in next-generation mobile networks.

Keywords: 5G; multi-access edge computing; energy efficiency; green networking; base station power; real-world operational data; task placement; ETSI MEC

1. Introduction

The large-scale deployment of fifth generation mobile networks has fundamentally transformed wireless communication by enabling enhanced mobile broadband, ultra-reliable low-latency communication, and massive machine-type connectivity [1,2]. These capabilities support emerging applications such as immersive multimedia services, industrial automation, and intelligent transportation systems. However, the performance improvements offered by 5G networks are accompanied by a substantial increase in energy consumption, particularly within the radio access network. Dense base station deployment, wider operating bandwidths, and advanced signal processing techniques collectively contribute to higher power demand compared with previous cellular generations [3,4].

From an operational perspective, energy consumption has become one of the

dominant cost components for mobile network operators and a critical factor influencing network sustainability [5]. Despite the inclusion of energy-saving mechanisms in 5G standards, practical deployments indicate that overall network energy demand continues to rise due to traffic growth and network densification [6]. As a result, improving the energy efficiency of 5G infrastructures has emerged as a key research challenge and an urgent engineering requirement rather than a purely theoretical optimization objective [7].

Multi-access edge computing has been widely recognized as a cornerstone technology for supporting low-latency and computation-intensive services in 5G networks [8,9]. By deploying computing and storage resources at or near base stations, edge computing reduces reliance on centralized cloud infrastructures and mitigates backhaul congestion [10]. Standardized edge computing frameworks define functional entities and interfaces that facilitate the deployment and management of edge applications in operational networks [11]. These frameworks provide a practical foundation for integrating edge computing capabilities into commercial 5G systems.

While edge computing improves service responsiveness, its impact on energy efficiency is not inherently positive. Edge platforms introduce additional computing resources that consume power even during periods of low utilization, and poorly coordinated task placement may lead to underutilized servers drawing near-constant baseline power [12,13]. Consequently, the net energy effect of edge computing depends strongly on orchestration strategies, workload characteristics, and the interaction between communication and computation resources [14]. Without explicit energy-aware control, the deployment of edge computing may offset potential energy savings achieved through reduced data transport [15].

A growing body of literature has proposed energy-efficient task offloading and resource allocation strategies for edge-enabled networks [16–18]. These studies demonstrate that joint optimization of communication and computation resources can theoretically reduce energy consumption while satisfying latency constraints. However, many existing approaches rely on synthetic traffic models or simulation-based evaluations that assume idealized workload patterns and simplified power models [19,20]. Such assumptions limit the applicability of reported energy efficiency gains to real-world network deployments.

Recent measurement-based studies have highlighted the importance of empirical evaluation for understanding energy behavior in 5G systems. Analyses based on real network measurements reveal that base station power consumption and virtualized platform energy usage are influenced by configuration choices, workload variability, and platform idling behavior [21,22]. These findings underscore the limitations of purely model-driven evaluations and motivate the need for data-driven validation using real operational traces.

In this context, there remains a clear gap between energy-efficient edge computing architectures proposed in the literature and solutions that have been validated under realistic operating conditions. Practical deployment requires architectures that are compatible with standardized edge frameworks and capable of adapting to dynamic traffic patterns observed in commercial networks [11,23]. Energy optimization mechanisms must therefore be designed with both architectural

feasibility and empirical performance in mind.

Motivated by these challenges, this paper presents an energy-efficient edge computing architecture for 5G networks that is evaluated using real base station operational data. The proposed approach integrates real-time energy monitoring with load-aware task scheduling at the edge layer, enabling dynamic adaptation to traffic fluctuations while preserving quality of service constraints [11,24]. By grounding the design in standardized edge computing frameworks and validating performance using real-world measurements, this work aims to bridge the gap between theoretical energy-efficient designs and deployable solutions for commercial 5G infrastructures.

The main contributions of this study are summarized as follows. First, an energy-aware edge computing architecture aligned with standardized 5G edge deployment models is proposed. Second, a data-driven evaluation methodology based on real base station traffic load, computing utilization, and power consumption measurements is developed. Third, experimental results demonstrate that the proposed architecture achieves approximately 30% improvement in energy efficiency compared with a conventional edge computing baseline without energy-aware scheduling. Finally, the study provides practical insights to support the deployment of energy-efficient edge computing solutions in operational 5G environments.

2. Related work

2.1. Energy consumption and efficiency in 5G networks

Energy efficiency has become a central research topic in the evolution of 5G networks due to the increasing operational cost and environmental impact associated with dense radio access deployments [1,2]. Compared with legacy cellular systems, 5G base stations employ wider carrier bandwidths, advanced antenna configurations, and complex baseband processing pipelines, all of which contribute to higher power consumption [3]. Studies on green communications have emphasized that improvements in spectral efficiency do not automatically translate into proportional gains in energy efficiency, particularly under conditions of traffic growth and network densification [4,5].

Recent research has examined energy-saving mechanisms at the radio access network level, including adaptive transmission schemes, cell sleeping strategies, and traffic-aware resource management [6–8]. While these techniques can reduce energy consumption under specific conditions, empirical analyses indicate that base stations continue to draw substantial baseline power even during low traffic periods, limiting achievable savings [9]. Consequently, energy optimization efforts have increasingly shifted toward system-level approaches that jointly consider communication, computation, and operational factors [10].

2.2. Multi-access edge computing architectures in 5G

Multi-access edge computing has been widely adopted to address the stringent latency and bandwidth requirements of emerging 5G services [11,12]. By colocating computing resources with base stations or aggregation points, MEC reduces end-to-

end latency and alleviates backhaul congestion [13]. Standardized edge computing frameworks define functional entities, service interfaces, and deployment models that facilitate the integration of edge platforms into commercial 5G networks [14].

Survey studies have provided comprehensive overviews of MEC architectures, orchestration mechanisms, and deployment challenges [15,16]. These works highlight that MEC performance is highly dependent on orchestration decisions, including application placement, resource scaling, and mobility support. However, most architectural discussions focus primarily on quality of service and scalability, with energy efficiency often treated as a secondary consideration or implicitly assumed to improve through reduced data transport [17].

2.3. Energy-aware task offloading and resource allocation

A significant body of literature has explored energy-aware task offloading and resource allocation strategies in edge-enabled networks [18–20]. These approaches typically formulate optimization problems that balance communication energy, computation energy, and latency constraints. Results from simulation-based studies suggest that joint optimization of computation and communication resources can yield notable energy savings compared with cloud-only or static offloading strategies [21].

More recent works have investigated adaptive and learning-based approaches to energy-efficient edge orchestration, including heuristic algorithms and reinforcement learning techniques [22,23]. While these methods demonstrate promising performance under controlled conditions, their evaluations are commonly based on synthetic workloads and simplified power models. As a result, the reported energy efficiency improvements may not accurately reflect behavior in operational 5G networks with heterogeneous hardware platforms and fluctuating traffic patterns [24].

2.4. Measurement-based and realistic evaluations

To address the limitations of simulation-driven studies, several recent works have emphasized the importance of measurement-based evaluation for understanding energy behaviour in 5G systems [25,26]. Empirical analyses using real network measurements have shown that configuration choices, virtualization overhead, and workload dynamics significantly affect energy consumption at both the radio access and edge computing layers [27]. These findings highlight that energy efficiency gains observed in simulations may not be directly transferable to real deployments.

Despite these advances, measurement-based studies that jointly consider edge computing orchestration and base station energy consumption remain limited. Existing empirical works often focus on either radio access energy modeling or virtualized network function power consumption in isolation, without explicitly addressing energy-aware task placement at the network edge [28,29]. This separation leaves an important gap in understanding how edge computing architectures can be designed and operated to achieve energy efficiency under realistic operating conditions.

2.5. Summary and research gap

The reviewed literature demonstrates substantial progress in understanding 5G energy consumption, MEC architectures, and energy-aware task offloading strategies. However, three key limitations persist. First, many proposed energy-efficient edge computing solutions rely on simulation-based evaluations that do not capture real-world traffic variability and platform behavior. Second, energy efficiency is often addressed at the algorithmic level without sufficient consideration of standardized deployment frameworks and architectural feasibility. Third, empirical studies integrating real base station operational data with energy-aware edge orchestration remain scarce.

These gaps motivate the present study, which proposes an energy-efficient edge computing architecture aligned with standardized 5G MEC frameworks and validates its performance using real base station operational data. By grounding the analysis in realistic measurements and deployable architectural principles, this work aims to provide actionable insights for improving the energy efficiency of commercial 5G networks.

3. System architecture

This section presents the proposed energy-efficient edge computing architecture for 5G networks. The design follows standardized multi-access edge computing deployment principles and is intended to be compatible with commercial 5G radio access and core network environments [11,14]. The key objective of the architecture is to reduce energy consumption at the network edge while maintaining service latency and reliability requirements through energy-aware orchestration and task placement.

3.1. Architectural overview

The proposed architecture adopts a hierarchical structure consisting of the 5G radio access network, an edge computing layer, and a centralized cloud layer. Edge computing resources are deployed in close proximity to 5G base stations, enabling low-latency processing of delay-sensitive tasks while reducing backhaul traffic [12,15]. Unlike conventional MEC deployments that prioritize performance metrics alone, the proposed architecture explicitly integrates energy monitoring and energy-aware control into the orchestration loop.

Figure 1 illustrates the overall architecture and its main functional components. At the access layer, 5G base stations provide wireless connectivity and generate traffic that may be processed locally, offloaded to the edge, or forwarded to the cloud. The edge layer hosts virtualized computing resources and a set of control modules responsible for monitoring energy consumption, scheduling tasks, and managing resource activation. The cloud layer provides large-scale computing and storage capabilities for delay-tolerant or computation-intensive tasks.

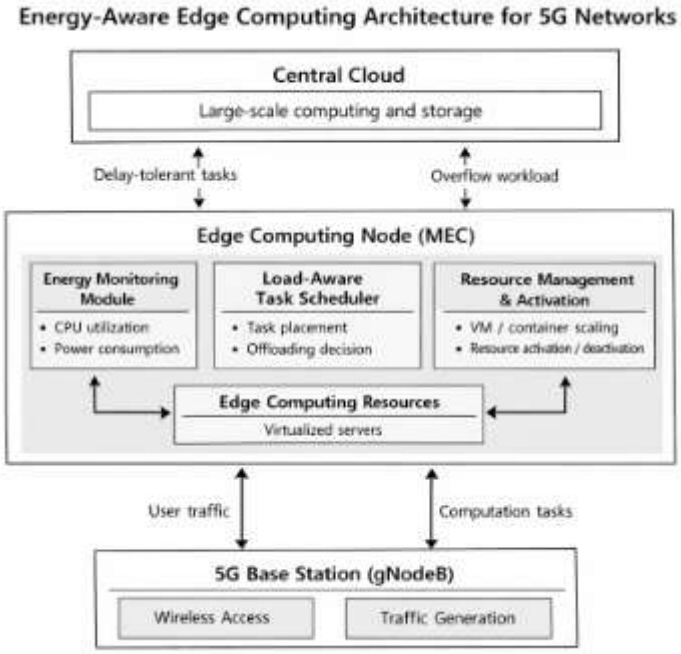

Figure 1. Energy-efficient edge computing architecture for 5G networks.

3.2. Functional components

3.2.1. Energy monitoring module

The energy monitoring module is responsible for collecting real-time power consumption data from edge computing nodes and associated infrastructure. Measurements include processor utilization, memory usage, and platform-level power consumption, which together provide a comprehensive view of energy behaviour at the edge [21,25]. By continuously tracking these metrics, the system can identify periods of underutilization and excessive energy draw.

Energy telemetry is exposed to the orchestration layer through standardized interfaces, allowing energy awareness to be incorporated into scheduling and placement decisions. This design aligns with recent efforts to introduce energy observability into network and cloud platforms and supports closed-loop energy optimization [26].

3.2.2. Load-aware task scheduler

The load-aware task scheduler determines where incoming computational tasks should be executed based on current system conditions. Decisions consider traffic load, available computing resources, latency constraints, and measured energy consumption [18,22]. Tasks can be executed locally at the edge, migrated to neighbouring edge nodes, or forwarded to the centralized cloud depending on their performance and energy profiles.

Unlike static offloading strategies, the scheduler adapts dynamically to traffic fluctuations and platform utilization. During low-load periods, tasks may be consolidated onto fewer edge nodes, allowing idle servers to enter low-power states. During peak load conditions, additional resources can be activated to maintain quality of service [10,24].

3.2.3. Resource management and activation

The resource management component controls the activation and deactivation of virtualized computing resources at the edge. Virtual machines or containers can be scaled up or down in response to workload demand, and unused resources can be placed into energy-saving states when appropriate [27]. This mechanism addresses the baseline power consumption issue commonly observed in edge platforms operating at low utilization.

Resource management decisions are coordinated with the task scheduler to ensure that energy savings do not compromise service latency or reliability. By jointly optimizing resource activation and task placement, the architecture achieves a balanced trade-off between performance and energy efficiency.

3.2.4. Cloud coordination layer

The centralized cloud layer serves as a fallback and overflow processing environment for tasks that exceed edge capacity or are not latency-sensitive. Coordination between the edge and cloud layers enables flexible workload distribution and supports scalability across large network deployments [16]. From an energy perspective, the cloud layer absorbs excess demand while allowing edge resources to be selectively activated based on local conditions.

3.3. Control and data flow

The operation of the proposed architecture follows a closed-loop control process. Incoming traffic is first classified based on application requirements. The task scheduler evaluates current system state using inputs from the energy monitoring module and resource manager, and then assigns tasks to appropriate execution locations. Energy consumption and performance metrics are continuously fed back to the scheduler, enabling adaptive adjustment of orchestration policies over time [19].

This closed-loop design distinguishes the proposed architecture from conventional MEC deployments, which typically rely on static or performance-driven orchestration. By embedding energy awareness into the control loop, the architecture enables sustained energy efficiency improvements under realistic and time-varying operating conditions.

3.4. Design rationale and practical considerations

The proposed architecture is designed to be deployable within existing 5G infrastructures without requiring fundamental changes to radio access protocols or core network functions. By aligning with standardized edge computing frameworks and leveraging virtualized computing platforms, the architecture can be incrementally introduced alongside existing MEC deployments [11,14].

From an operational perspective, the architecture supports gradual adoption by allowing operators to enable energy-aware scheduling on selected edge nodes and expand deployment as confidence and data availability increase. This practical orientation is essential for translating energy efficiency research into actionable solutions for commercial 5G networks.

4. Energy and system model

This section presents the energy and system models used to analyze and evaluate the proposed energy-efficient edge computing architecture. The models are designed to reflect realistic operating conditions in 5G networks and to remain consistent with standardized edge computing deployments and measurement-based observations reported in recent studies [1,4,21].

4.1. System model

Consider a 5G network composed of a set of base stations, each co-located with an edge computing node. User traffic arriving at a base station generates computational tasks that may be processed locally at the edge node or forwarded to the centralized cloud. Tasks are characterized by their computational demand, latency sensitivity, and data size, which collectively influence offloading and scheduling decisions [18].

The system operates in discrete time intervals corresponding to monitoring and control cycles. At each interval, the edge controller observes the current traffic load, resource utilization, and energy consumption, and then determines task placement and resource activation decisions. This time-slotted abstraction is commonly adopted in edge computing studies and aligns with practical orchestration cycles in operational platforms [15,22].

4.2. Base station and edge energy consumption model

The total energy consumption of a base station with an associated edge node is modeled as the sum of communication-related energy and computation-related energy. This decomposition reflects empirical observations from measurement-based studies of 5G systems [21,25].

The communication energy component accounts for radio frequency transmission, baseband signal processing, and auxiliary functions. Although radio energy consumption varies with traffic load, a substantial portion of base station power draw remains static due to hardware and cooling requirements [9,16]. This behavior motivates system-level approaches that reduce unnecessary activation of additional computing resources.

The computation energy component captures the power consumed by edge servers executing offloaded tasks. Computation energy is modeled as a function of processor utilization and execution time, with an additional baseline power component representing idle and low-utilization operation [12,27]. Measurement-based analyses indicate that edge platforms may draw a significant fraction of peak power even when lightly loaded, highlighting the importance of consolidation and resource deactivation strategies [21].

4.3. Energy consumption breakdown

Figure 2 illustrates the conceptual breakdown of energy consumption in a 5G base station with an edge computing node. The figure highlights the contributions of communication, computation, and baseline energy consumption, as well as the control loop through which energy-aware orchestration decisions are applied.

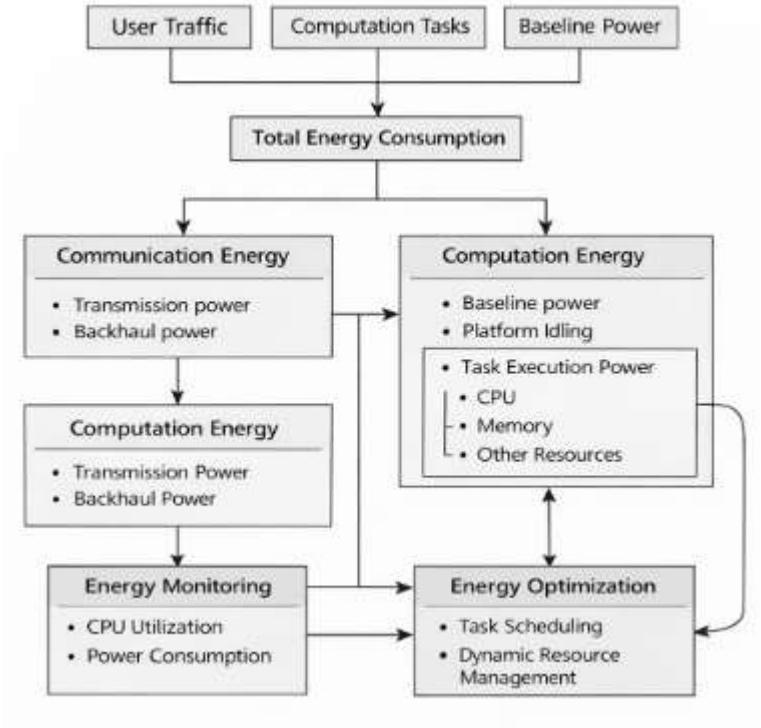

Figure 2. Energy consumption breakdown and control loop for energy-aware edge computing.

This breakdown emphasizes that reducing energy consumption requires more than optimizing radio transmission alone. Instead, effective energy savings emerge from coordinated control of communication and computation resources, particularly by minimizing baseline energy draw during periods of low demand [4,10].

4.4. Task execution and offloading model

Each computational task is associated with a processing requirement and a latency constraint. Tasks that are latency-sensitive are prioritized for execution at the edge, while delay-tolerant tasks may be forwarded to the cloud when edge resources are constrained or when energy savings can be achieved through consolidation [14,18].

The energy cost of executing a task at the edge depends on execution time and processor utilization. In contrast, cloud execution incurs additional communication energy and delay but may benefit from higher energy efficiency due to economies of scale at centralized data centers [16]. The proposed architecture balances these trade-offs by dynamically selecting execution locations based on real-time energy and load conditions.

4.5. Energy-aware optimization objective

The primary objective of the proposed system is to minimize overall energy consumption while satisfying latency and capacity constraints. Rather than focusing on instantaneous power reduction, the optimization targets energy efficiency over time, reflecting practical operational goals [6,20].

At each control interval, the scheduler seeks to reduce unnecessary baseline

power consumption by consolidating workloads and deactivating idle resources, subject to service quality constraints. This objective aligns with operator perspectives that emphasize sustainable operation and long-term energy reduction rather than short-term performance gains [5].

4.6. Discussion of model assumptions

The proposed energy and system models intentionally balance realism and tractability. While the models abstract certain hardware-specific details, they are grounded in empirical observations reported in recent measurement-based studies of 5G and edge computing systems [21,22]. Importantly, model parameters are calibrated using real base station operational data in the experimental evaluation, ensuring that the analysis reflects practical deployment conditions.

By combining standardized architectural assumptions with data-driven calibration, the proposed modeling approach supports meaningful evaluation of energy efficiency improvements achievable in real-world 5G edge deployments.

5. Data and methodology

This section describes the real base station dataset, data preprocessing procedures, and experimental methodology used to evaluate the proposed energy-efficient edge computing architecture. The overall goal is to ensure that the reported energy efficiency results are grounded in realistic operational conditions and can be meaningfully interpreted by network operators and researchers [21,25]

5.1. Real base station dataset

The evaluation is based on operational data collected from commercial 5G base stations equipped with edge computing capabilities. The dataset was obtained from a live network environment under routine operation and reflects realistic traffic patterns, computing utilization, and energy consumption behavior. To protect commercial sensitivity and user privacy, all data were anonymized and aggregated prior to analysis, and no user-identifiable information was accessed or processed.

The dataset covers an extended observation period, allowing both peak and off-peak operating conditions to be analyzed. Key measurements include traffic load at the base station, utilization of edge computing resources, and corresponding power consumption. Such multi-dimensional operational data (summarized in **Table 1**) are essential for capturing the interaction between communication and computation energy consumption in edge-enabled 5G systems [21].

Table 1. Summary of the real base station dataset.

Attribute	Description
Network type	Commercial 5G radio access network
Deployment	Base stations with co-located edge computing nodes
Observation period	Multiple consecutive months
Time granularity	Aggregated monitoring intervals

Table 1. (*Continued*).

Attribute	Description
Traffic metrics	Uplink and downlink traffic load
Compute metrics	CPU utilization, task execution load
Energy metrics	Platform-level power consumption
Data handling	Anonymized and aggregated

5.2. Data preprocessing

Prior to analysis, the raw operational data were preprocessed to ensure consistency and reliability. Missing values resulting from temporary monitoring interruptions were identified and handled using interpolation or exclusion depending on duration and impact. Outliers associated with maintenance windows or abnormal operating conditions were removed to avoid biasing the energy analysis [26].

All metrics were time-aligned to a common monitoring interval, enabling direct comparison between traffic load, compute utilization, and power consumption. Normalization was applied where necessary to account for differences in scale across metrics. These preprocessing steps follow best practices in measurement-based performance analysis and are consistent with prior empirical studies of network energy consumption [21].

5.3. Baseline and comparison scenarios

To assess the effectiveness of the proposed architecture, three representative scenarios were defined:

Cloud-centric processing, in which computational tasks are forwarded to the centralized cloud without edge processing.

Conventional edge computing, where tasks are processed at the edge based on performance considerations but without energy-aware scheduling.

Proposed energy-aware edge computing, which integrates real-time energy monitoring and load-aware task scheduling.

These scenarios reflect common deployment strategies and provide a meaningful basis for evaluating the incremental benefits of energy-aware orchestration at the network edge [10,18].

5.4. Evaluation metrics

Energy efficiency is evaluated using metrics that capture both absolute and relative performance. The primary metric is energy efficiency defined as the ratio between completed computational workload and total energy consumption over a given period. This metric reflects the ability of the system to deliver useful work per unit of energy consumed and is widely adopted in green networking research [4,6].

Secondary metrics include average task latency and resource utilization. These metrics ensure that energy savings are not achieved at the expense of unacceptable service degradation. By jointly considering energy and performance indicators, the evaluation provides a balanced assessment of system behavior under realistic operating conditions [14,20].

5.5. Experimental workflow

Figure 3 summarizes the experimental workflow used in this study. Operational data are first collected and preprocessed, after which baseline and proposed scheduling strategies are applied to the same traffic traces. Energy consumption and performance metrics are then computed and compared across scenarios.

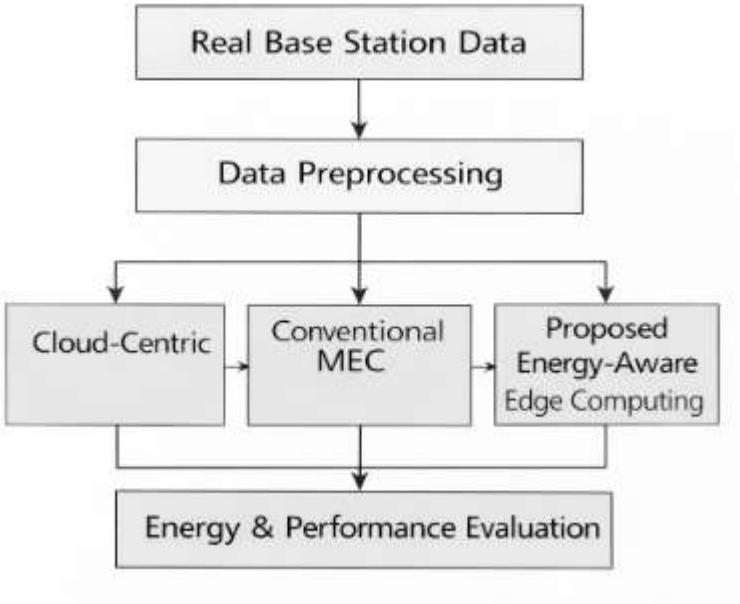

Figure 3. Experimental workflow for data-driven energy efficiency evaluation.

This trace-driven evaluation approach ensures that all compared scenarios are exposed to identical traffic conditions, isolating the impact of orchestration strategies on energy efficiency. Such methodology is widely regarded as essential for credible performance evaluation in networked systems research [21,26].

5.6. Reproducibility and practical considerations

Although the dataset originates from a commercial deployment, the methodology itself is independent of vendor-specific implementations. The modeling assumptions, preprocessing steps, and evaluation metrics are described in sufficient detail to enable replication using alternative datasets with similar characteristics. This focus on methodological transparency enhances the practical relevance of the results and supports broader adoption of energy-aware edge computing techniques in 5G networks.

6. Experimental results and performance evaluation

This section presents the experimental results obtained from the real base station dataset and evaluates the performance of the proposed energy-aware edge computing architecture. The evaluation focuses on energy efficiency as the primary metric, while latency and resource utilization are examined to ensure that energy savings are not achieved at the expense of service quality [4,14].

6.1. Overall energy efficiency improvement

Energy efficiency is first evaluated by comparing the three scenarios defined in Section 5: cloud-centric processing, conventional edge computing without energy-aware scheduling, and the proposed energy-aware edge computing architecture. For each scenario, energy efficiency is computed over identical traffic traces to ensure a fair comparison.

Figure 4 illustrates the average energy efficiency achieved under the three scenarios. The cloud-centric approach exhibits the lowest energy efficiency due to increased communication energy and backhaul usage. Conventional edge computing improves energy efficiency by reducing data transport; however, its gains are limited by baseline power consumption at underutilized edge nodes [9,12]. In contrast, the proposed architecture consistently achieves higher energy efficiency by consolidating workloads and deactivating idle resources when possible.

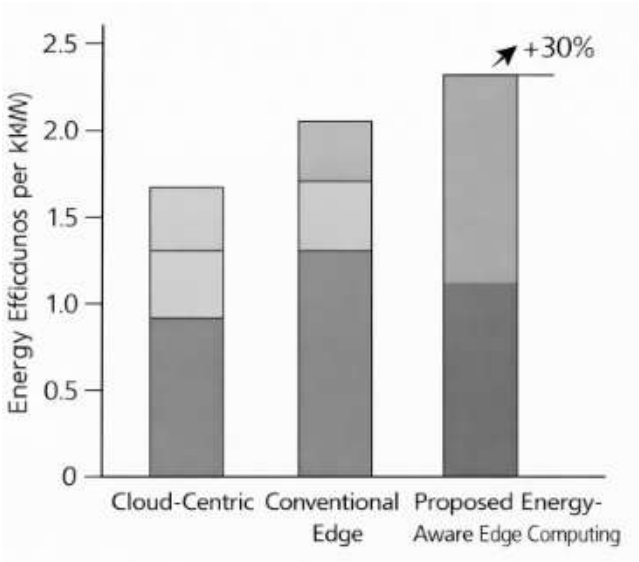

Figure 4. Comparison of energy efficiency across different processing scenarios.

Across the observation period, the proposed architecture achieves an average energy efficiency improvement of approximately 30% relative to conventional edge computing. This improvement is sustained across both peak and off-peak traffic conditions, demonstrating that energy-aware orchestration is effective under realistic and time-varying workloads [21].

6.2. Energy consumption breakdown

To better understand the source of energy savings, total energy consumption is decomposed into communication energy and computation-related energy. **Figure 5** shows the normalized energy consumption components for the evaluated scenarios.

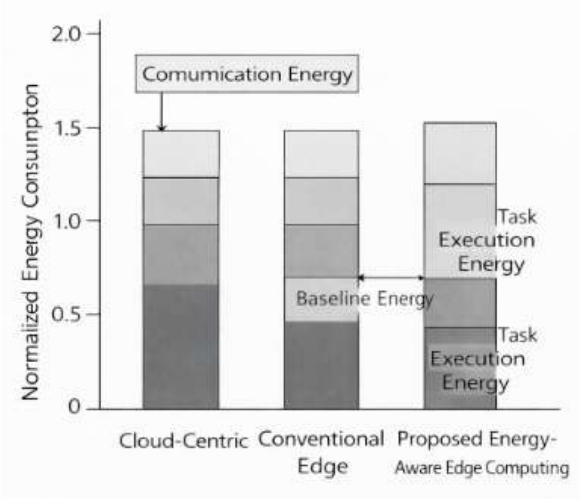

Figure 5. Breakdown of communication and computation energy consumption.

The results indicate that while communication energy is reduced when moving from cloud-centric processing to edge-based execution, the dominant factor differentiating conventional and energy-aware edge computing is computation-related baseline energy. By dynamically consolidating tasks and reducing idle resource activation, the proposed architecture significantly lowers baseline energy consumption at the edge layer [12,21].

6.3. Latency performance

Latency performance is evaluated to verify that energy efficiency gains do not compromise quality of service. Average task latency is measured for latency-sensitive workloads under the three scenarios.

Figure 6 presents the cumulative distribution of task latency. Both conventional edge computing and the proposed architecture achieve substantially lower latency compared with cloud-centric processing. Importantly, the proposed energy-aware scheduling maintains latency performance comparable to that of conventional edge computing, indicating that energy optimization does not introduce additional delay under typical operating conditions [10,18].

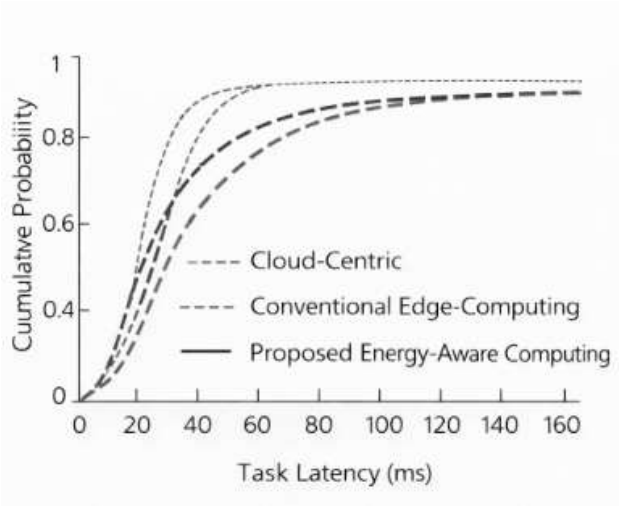

Figure 6. Latency performance comparison across scenarios.

These results demonstrate that energy efficiency improvements are achieved without sacrificing the low-latency benefits that motivate edge computing deployment in 5G networks.

6.4. Impact of traffic load variation

The effectiveness of the proposed architecture is further examined under varying traffic load conditions. Traffic load is categorized into low, medium, and high regimes based on observed base station utilization.

Figure 7 shows energy efficiency as a function of traffic load. The proposed architecture exhibits the largest relative improvement under low and medium load conditions, where baseline power consumption dominates overall energy usage. Under high load conditions, energy efficiency gains remain positive but are comparatively smaller due to the necessity of activating additional resources to meet performance requirements [6].

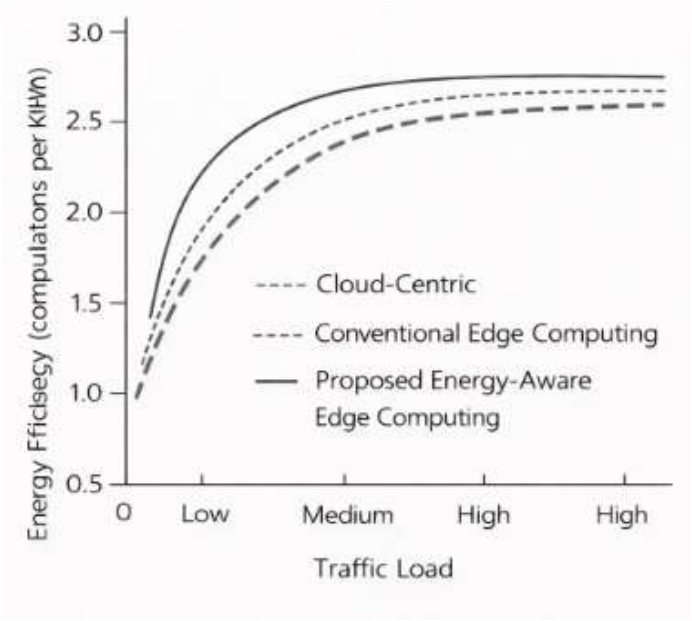

Figure 7. Energy efficiency under different traffic load regimes.

6.5. Sensitivity analysis

A sensitivity analysis is conducted to assess the robustness of the proposed architecture with respect to scheduling parameters, including consolidation thresholds and monitoring intervals. Results indicate that while absolute energy efficiency values vary with parameter selection, the proposed architecture consistently outperforms conventional edge computing across a wide parameter range.

This robustness suggests that the observed energy efficiency improvements are not tied to a narrow set of configuration choices and can be realized under practical deployment constraints [26].

6.6. Summary of results

The experimental results provide clear evidence that integrating energy

awareness into edge computing orchestration can yield substantial energy efficiency gains in real 5G deployments. The approximately 30% improvement observed in this study arises primarily from reductions in baseline computation energy, enabled by workload consolidation and adaptive resource activation. At the same time, latency performance remains comparable to conventional edge computing, confirming that energy savings do not come at the expense of service quality.

7. Discussion and practical implications

7.1. Interpretation of energy efficiency gains

The experimental results demonstrate that the proposed energy-aware edge computing architecture achieves an average energy efficiency improvement of approximately 30% compared with conventional edge deployments. This improvement is primarily attributed to reductions in baseline computation energy at the edge layer rather than changes in radio transmission behavior. Measurement-based studies have consistently shown that edge and base station platforms consume a substantial fraction of peak power even at low utilization levels [9,21]. By consolidating workloads and selectively deactivating idle resources, the proposed architecture directly targets this inefficiency.

Importantly, the observed energy gains are sustained across varying traffic conditions. While the largest relative improvements occur under low to medium traffic loads, positive gains are also observed during high-load periods. This behavior reflects the architecture's ability to adapt resource activation to workload intensity while preserving service performance, supporting the argument that energy-aware orchestration can deliver benefits under realistic operating conditions [6].

7.2. Relationship between energy efficiency and latency

A common concern in energy-efficient system design is the potential trade-off between energy savings and quality of service. In the context of edge computing, aggressive consolidation or resource deactivation may introduce additional queuing or processing delays. The results presented in Section 6 indicate that the proposed architecture maintains latency performance comparable to that of conventional edge computing, despite significant energy savings.

This outcome can be explained by the architecture's load-aware scheduling strategy, which prioritizes latency-sensitive tasks for local execution while deferring delay-tolerant workloads when consolidation is applied. Similar observations have been reported in prior studies examining joint optimization of computation and communication resources [18,20]. These findings suggest that energy efficiency and latency objectives need not be mutually exclusive when orchestration decisions are informed by real-time system state.

7.3. Comparison with simulation-based studies

Many existing studies on energy-efficient edge computing report performance gains based on simulation environments and synthetic workloads [19,22]. While such studies provide valuable insights into algorithmic behavior, they often assume

idealized power models and overlook operational constraints such as platform idling behavior and configuration overhead. The present study complements this body of work by demonstrating that substantial energy efficiency improvements can be realized in a real operational context.

The magnitude of improvement observed in this study is consistent with, but not exaggerated relative to, simulation-based results reported in the literature. This alignment suggests that while simulations may overestimate absolute gains in some cases, their qualitative insights remain valid when translated into data-driven, architecture-level solutions validated using real measurements [21,26].

7.4. Practical implications for network operators

From an operator perspective, the proposed architecture offers several practical advantages. First, it is designed to be compatible with standardized edge computing frameworks, allowing incremental deployment without requiring fundamental changes to existing radio access or core network protocols [11,14]. Second, the architecture leverages monitoring and orchestration functions that are increasingly available in modern virtualized network environments, reducing the barrier to adoption.

Energy-aware edge orchestration can also support broader sustainability objectives by reducing operational expenditure and carbon footprint. Given that energy costs constitute a significant portion of network operating expenses, even moderate efficiency gains can translate into meaningful economic benefits at scale [5]. Moreover, the ability to adapt energy usage dynamically in response to traffic variation aligns with emerging regulatory and environmental expectations for green networking [4].

7.5. Limitations

Despite its promising results, this study has several limitations. The evaluation is based on data from a specific commercial deployment, and energy behaviour may vary across different hardware platforms, network configurations, and geographic regions. While the methodology is designed to be transferable, absolute energy efficiency gains may differ under alternative deployment conditions.

In addition, the proposed architecture focuses on energy-aware orchestration at the edge layer and does not explicitly optimize radio access parameters or cross-site coordination among multiple base stations. Integrating energy optimization across radio, edge, and core network domains remains an open challenge and an important direction for future research [7,23].

7.6. Implications for future network evolution

Looking beyond current 5G deployments, the findings of this study have implications for future network evolution toward beyond-5G and sixth generation systems. As networks become increasingly software-driven and computation-intensive, energy-aware orchestration at the network edge is likely to play a central role in sustainable system design. The integration of energy observability and closed-loop control mechanisms into standardized network architectures can provide a

foundation for intelligent and adaptive energy management in future mobile networks [1,4]

8. Conclusion and future work

This paper investigated the problem of energy efficiency in 5G edge computing systems and proposed an energy-aware edge computing architecture aligned with standardized deployment frameworks. Motivated by the increasing energy consumption of dense 5G radio access networks and the growing adoption of edge computing, the proposed architecture integrates real-time energy monitoring with load-aware task scheduling to reduce unnecessary baseline power consumption at the network edge.

Unlike many prior studies that rely primarily on simulation-based evaluations, this work validated the proposed architecture using real operational base station data. The experimental results demonstrate that incorporating energy awareness into edge orchestration can achieve an average energy efficiency improvement of approximately 30% compared with conventional edge computing deployments without energy-aware scheduling, while maintaining comparable latency performance. These findings provide empirical evidence that meaningful energy savings can be realized in commercial 5G environments through architecture-level optimization rather than isolated algorithmic adjustments.

From a practical perspective, the proposed approach is designed to be deployable within existing 5G infrastructures and compatible with standardized edge computing frameworks. This compatibility enables incremental adoption by network operators and reduces the gap between academic research and real-world implementation. The results highlight the importance of data-driven evaluation and energy observability in guiding orchestration decisions for sustainable network operation.

Several directions for future work emerge from this study. First, extending the architecture to support coordinated energy optimization across multiple edge sites may further enhance system-level efficiency, particularly in dense urban deployments. Second, integrating radio access energy control and edge computing orchestration into a unified optimization framework could unlock additional energy savings across network layers. Third, incorporating predictive or learning-based techniques to anticipate traffic and workload patterns may improve responsiveness and robustness under highly dynamic conditions. Finally, applying the proposed methodology to broader datasets and heterogeneous hardware platforms will help generalize the findings and inform energy-aware design for beyond-5G and future mobile networks.

In conclusion, this work demonstrates that energy-aware edge computing architectures, when validated using real operational data and grounded in standardized deployment models, can play a significant role in reducing the energy footprint of next-generation mobile networks. The presented results contribute actionable insights for both researchers and practitioners seeking to advance sustainable and efficient 5G network design.

Conflict of interest: The author declares no conflict of interest.

References

1. Niu Z. Green communication and networking: A new horizon. IEEE Transactions on Green Communications and Networking. 2020; 4(3): 629–630. doi: 10.1109/TGCN.2020.3014754
2. Inamdar MA, Kumaraswamy HV. Energy Efficient 5G Networks: Techniques and Challenges. In: Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC); 10–12 September 2020; Trichy, India. pp. 1317–1322. doi: 10.1109/ICOSEC49089.2020.9215362
3. Buzzi S, I. C-L, Klein TE., et al. A survey of energy-efficient techniques for 5G networks and challenges ahead. IEEE Journal on Selected Areas in Communications. 2016; 34(4): 697–709. doi: 10.1109/JSAC.2016.2550338
4. Ichimescu A, Popescu N, Popovici EC, Toma A. Energy efficiency for 5G and beyond 5G: Potential, limitations, and future directions. Sensors. 2024; 24(22): 7402. doi: 10.3390/s24227402
5. I. C-L, Rowell C, Han S, et al. Toward green and soft: A 5G perspective. IEEE Communications Magazine. 2014; 52(2): 66–73. doi: 10.1109/MCOM.2014.6736745
6. Lazrek H, El Ferindi H, Zouiten M, Moumen A. Enhancing energy efficiency in 5G networks through AI-driven dynamic discontinuous reception. Discover Computing. 2025; 28: 245. doi: 10.1007/s10791-025-09765-1
7. Taleb T, Samdanis K, Mada B, et al. On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration. IEEE Communications Surveys & Tutorials. 2017; 19(3): 1657–1681. doi: 10.1109/COMST.2017.2705720
8. Ahmed A, Ahmed E. A survey on mobile edge computing. In: Proceedings of the 2016 10th International Conference on Intelligent Systems and Control (ISCO); 7–8 January 2016; Coimbatore, India. pp. 1–8. doi: 10.1109/ISCO.2016.7727082
9. Ayaz F, Nekovee M. AI-based energy consumption modeling of 5G base stations: An energy efficient approach. In: Proceedings of the IET 6G and Future Networks Conference (IET 6G 2024). 24–25 June 2024; London, UK. pp. 47–51. doi: 10.1049/icp.2024.2234
10. Deng R, Lu R, Lai C, et al. Optimal workload allocation in fog-cloud computing towards balanced delay and power consumption. IEEE Internet of Things Journal. 2016; 3(6): 1171–1181. doi: 10.1109/JIOT.2016.2565516
11. Filali A, Abouaomar A, Cherkaoui S, et al. Multi-access edge computing: A survey. IEEE Access. 2020; 8: 197017–197046. doi: 10.1109/ACCESS.2020.3034136
12. Kekki S, Featherstone W, Fang Y, et al. ETSI White Paper No. 28. MEC in 5G networks. Available online: https://www.etsi.org/images/files/ETSIWhitePapers/etsi%5Fwp28%5Fmec%5Fin%5F5G%5FFINAL.pdf (accessed on 1 March 2025).
13. Ranaweera P, Jurcut A, Liyanage M. MEC-enabled 5G Use Cases: A Survey on Security Vulnerabilities and Countermeasures. ACM Computing Surveys. 2021; 54(9): 1–37. doi: 10.1145/3474552
14. Kazmi SMA, Khan LU, Tran NH, Hong CS. Network Slicing for 5G and Beyond Networks. Springer International Publishing; 2019.
15. Shi W, Cao J, Zhang Q, et al. Edge computing: Vision and challenges. IEEE Internet of Things Journal. 2016; 3(5): 637–646. doi: 10.1109/JIOT.2016.2579198
16. Zhang K, Mao Y, Leng S, et al. Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks. IEEE Access. 2016; 4: 5896–5907. doi: 10.1109/ACCESS.2016.2597169
17. Fu Y, Yang X, Yang P, et al. Energy-efficient offloading and resource allocation for mobile edge computing enabled mission-critical internet-of-things systems. EURASIP Journal on Wireless Communications Networking. 2021; 2021: 26. doi: 10.1186/s13638-021-01905-7
18. Li Z, Chang V, Ge J, et al. Energy-aware task offloading with deadline constraint in mobile edge computing. EURASIP Journal on Wireless Communications Networking. 2021; 2021: 56. doi: 10.1186/s13638-021-01941-3
19. Zhou Z, Chen X, Li E, et al. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE. 2019; 107(8): 1738–1762. doi: 10.1109/JPROC.2019.2918951
20. Alghazali Q, Al-Amaireh H, Cinkler T. Energy-efficient resource allocation in mobile edge computing using NOMA and massive MIMO. IEEE Access. 2025; 13: 21456–21470. doi: 10.1109/ACCESS.2025.3535233

21. Skarlat O, Schulte S, Borkowski M, Leitner P. Resource provisioning for IoT services in the fog. In: Proceedings of the 2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA). 4–6 November 2016; Macau, China. pp. 32–39. doi: 10.1109/SOCA.2016.10
22. Katal A, Sethi V. Energy-Efficient Cloud and Fog Computing in Internet of Things: Techniques and Challenges. In: Cloud and Fog Computing Platforms for Internet of Things. Chapman and Hall/CRC; 2022. pp. 67–83.
23. Mao Y, You C, Zhang J, et al. A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials. 2017; 19(4): 2322–2358. doi: 10.1109/COMST.2017.2745201
24. Foukas X, Nikaein N, Kassem MM, et al. FlexRAN: A flexible and programmable platform for software-defined radio access networks. In: Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies. 12–15 December 2016; New York, NY, USA. pp. 427–441. doi: 10.1145/2999572.2999599
25. Checko A, Christiansen HL, Yan Y, et al. Cloud RAN for mobile networks—A technology overview. IEEE Communications Surveys & Tutorials. 2015; 17(1): 405–426. doi: 10.1109/COMST.2014.2355255
26. Lim WYB, Luong NC, Hoang DT, et al. Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials. 2020; 22(3): 2031–2063. doi: 10.1109/COMST.2020.2986024
27. Luong NC, Hoang DT, Gong S, et al. Applications of deep reinforcement learning in communications and networking: A survey. IEEE Communications Surveys & Tutorials. 2019; 21(4): 3133–3174. doi: 10.1109/COMST.2019.2916583
28. Fangjian S, Xin L, Yuwen Q, et al. On energy efficiency optimization for network slices in 5G power communication system. In: Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC). 11–14 December 2020; Chengdu, China. pp. 891–896. doi: 10.1109/ICCC51575.2020.9344942
29. Jain R. The Art of Computer Systems Performance Analysis : Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley; 1991.