SOCIAL AND MANAGEMENT SCIENCES RESEARCH JOURNAL

www.smsrj.com

**p-ISSN:** 2791-0415 **e-ISSN:** 2791-0423

**Volume:4 Issue: 02(2026)**

# Joint Computing, Communication, and Interface Resource Allocation for Heterogeneous Edge Systems

***[1]Brandon C. Khan, [2]Zoe K. Hernandez, [3]Joshua Roberts***

*[1,3]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Champaign, IL 61801, USA*

*[2]Industrial and Systems Engineering, College of Engineering, Columbus Campus, The Ohio State University, Columbus, OH 43210, USA*

**Abstract:**

*The proliferation of ubiquitous internet of things devices and computationally intensive applications has fundamentally transformed the landscape of modern wireless networks. Heterogeneous edge computing has emerged as a promising paradigm to alleviate the burden on centralized cloud infrastructures by migrating computational resources to the network periphery. However, the extreme heterogeneity in device capabilities, varying communication channels, and diverse network interfaces poses significant challenges for efficient resource orchestration. This paper presents a comprehensive investigation into the joint allocation of computing, communication, and interface resources in heterogeneous edge systems. We formulate a multi-objective optimization problem aimed at minimizing the system wide computational delay and energy consumption while accommodating the strict constraints of multidimensional resource capacities. Given the highly non-convex nature of the formulated problem, we propose an iterative optimization framework grounded in block coordinate descent and successive convex approximation techniques. By decoupling the original complex problem into tractable subproblems, the proposed algorithmic methodology achieves near optimal performance with manageable computational complexity. Extensive theoretical analyses and simulated experiments demonstrate that the proposed joint allocation strategy significantly outperforms conventional isolated resource management schemes, providing a robust, scalable, and highly efficient solution for next generation edge computing environments.*

***Keywords:*** *Edge Computing, Resource Allocation, Optimization Theory, Heterogeneous Networks*

## 1. Introduction

### 1.1 The Paradigm of Edge Computing

The rapid development of wireless communication technologies and the exponential proliferation of smart mobile devices have catalyzed the emergence of computationally

demanding applications. Technologies such as augmented reality, autonomous vehicular navigation, and real time industrial automation require massive computational power and ultra low latency communication. Traditional cloud computing architectures, which rely on centralized remote servers to process data, are increasingly inadequate for these emerging applications due to the inevitable propagation delays and heavy backhaul network congestion. To address these critical bottlenecks, mobile edge computing has been proposed as a transformative network architecture. By deploying edge servers equipped with substantial storage and computing capabilities at the network edge, such as at base stations or wireless access points, edge computing allows mobile devices to offload their heavy computational tasks to nearby infrastructure. This fundamental paradigm shift significantly reduces service latency, conserves device battery life, and alleviates the traffic burden on the core network. Initial studies in this domain have primarily focused on the feasibility of offloading single tasks to homogeneous edge servers [1]. However, as the technological landscape matures, it is becoming increasingly evident that modern edge computing environments are characterized by massive scale and profound heterogeneity, requiring more sophisticated management strategies. The integration of edge computing into the standard networking architecture necessitates a complete reevaluation of how resources are provisioned, scheduled, and maintained across diverse geographical regions and temporal variations. Research has shown that without intelligent orchestration, edge servers can easily become overwhelmed, leading to degraded performance that negates the very benefits the architecture was designed to provide [2].

**1.2 Heterogeneity in Edge Systems**

The contemporary edge computing ecosystem is inherently heterogeneous across multiple dimensions. First, the end user devices range from simple environmental sensors with minimal processing capabilities and severe energy constraints to sophisticated smartphones and autonomous vehicles equipped with multi core processors and large battery capacities. Second, the edge infrastructure itself comprises varying tiers of computational nodes, including micro data centers, roadside units, and small cell base stations, each possessing distinct computational capacities, memory limits, and energy profiles. Third, the communication environment is characterized by multiple overlapping radio access technologies, including fifth generation cellular networks, dedicated short range communications, and various local area wireless standards. This profound heterogeneity complicates the task offloading decision process. A homogeneous network model, which assumes uniform capacities and identical communication links, is entirely inadequate for capturing the intricacies of real world deployments. Consequently, recent literature has emphasized the critical need to model edge systems with high fidelity, explicitly accounting for the diverse characteristics of both the requesting devices and the serving infrastructure [3]. The heterogeneity also extends to the types of computational tasks generated by user devices. Some tasks are highly parallelizable and data intensive, while others are strictly sequential and require specialized hardware accelerators. Managing these diverse tasks across heterogeneous nodes demands a dynamic and highly adaptable resource allocation framework that can intelligently match task requirements with available system resources in real time.

**1.3 The Need for Joint Resource Allocation**

Historically, network resource management and computational resource allocation have been treated as independent optimization problems. Network engineers focused on maximizing spectral efficiency and minimizing interference, while software and systems engineers focused on scheduling central processing unit cycles and managing memory architectures. However, in heterogeneous edge systems, computation and communication are inextricably linked. The process of task offloading intrinsically involves transmitting data over a wireless communication

channel, processing the data at an edge server, and potentially returning the results to the mobile device. Optimizing communication resources such as transmit power and bandwidth without considering the computational load on the target server may result in a situation where data arrives quickly but experiences massive queuing delays at the processor. Conversely, optimizing computing resource allocation without regard for channel conditions may lead to severe transmission bottlenecks, rendering the computational efficiency irrelevant. Furthermore, modern devices are increasingly equipped with multiple network interfaces, allowing them to simultaneously connect to different access points using varied communication protocols. The selection of which interface to use dictates the communication parameters and significantly impacts energy consumption and delay. Therefore, treating computing, communication, and interface selection as isolated problems leads to highly suboptimal system performance. Joint resource allocation is imperative to fully harness the potential of heterogeneous edge systems, ensuring that multi dimensional resources are orchestrated synergistically to achieve global system optimization.

### 1.4 Contributions and Organization

This paper aims to bridge the existing gap in edge network orchestration by presenting a holistic and rigorous framework for the joint allocation of computing, communication, and interface resources. The primary objective is to formulate a comprehensive system model that accurately reflects the extreme heterogeneity of modern edge networks, incorporating variable device capabilities, multi tier edge servers, and diverse communication interfaces. We define a multi objective optimization problem that seeks to minimize the weighted sum of energy consumption and execution delay across all mobile devices in the network, subject to strict capacity constraints on computing power, transmission energy, and bandwidth limits. Given that the resulting mathematical formulation is a mixed integer non linear programming problem, which is well known to be non deterministic polynomial time hard, finding a global optimum is computationally prohibitive for real time applications. Therefore, we propose an innovative iterative algorithm based on block coordinate descent and successive convex approximation to decompose the complex global problem into manageable subproblems. These subproblems are solved sequentially to achieve a highly efficient, near optimal solution. The rest of this paper is systematically organized as follows. Section 2 provides a detailed background and literature review of existing resource allocation strategies in edge computing. Section 3 introduces the comprehensive system architecture and rigorously formulates the mathematical models for communication, computation, and interface mechanics. Section 4 elaborates on the proposed joint resource allocation methodology, detailing the mathematical transformations and algorithmic framework. Section 5 presents extensive experimental results, discusses the performance of the proposed method against conventional baselines, and concludes the paper with directions for future research.

## 2. Background and Related Work

### 2.1 Evolution of Mobile Edge Computing

The conceptual foundation of edge computing was initially proposed to overcome the physical limitations of cloud computing, specifically the high latency caused by the geographical distance between end users and central data centers. Early implementations of edge computing were relatively straightforward, focusing on deploying localized servers at macro base stations to cache popular content and perform basic data processing tasks [4]. These initial architectures demonstrated significant reductions in service latency and backhaul traffic. However, as the volume of internet of things devices exploded, the architecture rapidly evolved into a multi tiered, distributed computing paradigm. Researchers began exploring the theoretical limits of edge computing, focusing on queuing models and deterministic task scheduling to understand

the fundamental tradeoffs between local execution and remote offloading [5]. Subsequent studies expanded the scope to include energy harvesting devices, vehicular networks, and unmanned aerial vehicle assisted edge computing [6]. This evolutionary trajectory highlights a continuous shift from simple, static offloading policies to highly dynamic, context aware orchestration mechanisms. Current state of the art architectures envision a highly fluid compute continuum, where resources are dynamically pooled and shared across devices, edge nodes, and regional clouds [7]. Despite these advancements, the immense complexity introduced by architectural heterogeneity remains a formidable challenge, prompting extensive research into advanced mathematical optimization and machine learning techniques to manage the decentralized infrastructure.

**2.2 Computing Resource Allocation Strategies**

The allocation of computational resources, primarily defined in terms of processor clock frequencies or virtual machine instances, has been a central focus of edge computing research. A vast majority of early literature adopted game theoretic approaches to model the interactions between selfish mobile devices competing for limited edge server capacities. For instance, non cooperative games have been extensively utilized to achieve Nash equilibrium states where devices autonomously decide their offloading strategies based on pricing mechanisms enforced by the edge service providers [8]. Other researchers have formulated the computing resource allocation problem as a mixed integer programming model, applying heuristic algorithms such as particle swarm optimization and genetic algorithms to find acceptable solutions within reasonable timeframes [9]. In highly dynamic environments where task arrival rates and sizes fluctuate unpredictably, Lyapunov optimization techniques have been widely adopted to ensure long term system stability while minimizing time averaged delays [10]. Furthermore, with the advent of hardware virtualization, containerized microservices have introduced a new layer of complexity to computing allocation. Modern allocation schemes must now account for container startup times, memory isolation overheads, and the intricacies of orchestrating microservices across multiple physical nodes [11]. While these computational allocation strategies have proven effective in isolated scenarios, their failure to simultaneously optimize the underlying wireless transmission parameters often results in significant performance bottlenecks in practical, real world deployments.

**2.3 Communication Resource Management**

Parallel to the developments in computing allocation, communication resource management in wireless networks has a rich history of theoretical and practical advancements. In the context of edge offloading, communication resources typically encompass transmit power, channel bandwidth, and the assignment of subcarriers or resource blocks. The fundamental objective is to maximize the data transmission rate while minimizing energy consumption and mitigating co channel interference among multiple transmitting devices [12]. Traditional approaches have heavily relied on fractional programming and water filling algorithms to optimally distribute power across available frequency bands [13]. With the integration of multiple input multiple output antenna technologies, spatial multiplexing and beamforming have become critical components of communication resource allocation, allowing multiple devices to offload tasks simultaneously over the same frequency spectrum without destructive interference. Recent studies have also explored the use of non orthogonal multiple access techniques in edge computing, which permit multiple users to share the same time frequency resource block by separating their signals in the power domain [14]. While these sophisticated communication strategies significantly enhance the uplink and downlink data rates, optimizing them in a vacuum without considering the computational capabilities of the receiving edge server is inherently flawed. If an edge server lacks the processor capacity to handle a sudden influx of data optimized

by an advanced multiple access scheme, the resulting computational queuing delay will entirely negate the transmission time savings.

### 2.4 Interface and Multi-Access Resource Orchestration

The integration of multiple communication interfaces on modern mobile devices has introduced a novel dimension to resource allocation. Devices are now routinely equipped with cellular transceivers, wireless local area network adapters, and short range communication modules. Each interface presents a unique profile regarding bandwidth, transmission power requirements, reliability, and monetary cost [15]. The decision of which interface, or combination of interfaces, to utilize for task offloading is a non trivial problem. Initial research in multi access edge computing proposed simple threshold based heuristic rules for interface selection, typically prioritizing wireless local area networks due to their low cost and switching to cellular networks only when the local connection degraded. More sophisticated approaches have modeled the interface selection problem using Markov decision processes, where the device learns the optimal selection policy over time based on historical channel conditions and network congestion metrics [16]. However, dynamic interface selection profoundly impacts both the communication delay and the subsequent computing resource allocation at the edge server, as different interfaces may route traffic to entirely different tiers of the edge infrastructure. Jointly optimizing interface selection along with transmit power, bandwidth, and computational frequency remains an open research problem. Existing frameworks often simplify one or more of these dimensions to maintain mathematical tractability, thereby compromising the overall system efficiency [17]. This paper directly addresses this gap by presenting a fully integrated optimization framework that rigorously coordinates all three resource dimensions without oversimplifying the underlying physical models.

## 3. System Architecture and Mathematical Modeling

### 3.1 Heterogeneous Edge Network Topology

We consider a highly heterogeneous edge computing system operating within a defined geographical region. The network architecture comprises a set of multi tiered edge servers, denoted by the index set, which are collocated with various wireless access points, including macro cellular base stations, small cell routers, and dedicated roadside units. These edge servers possess heterogeneous computational capacities, meaning their maximum processor clock frequencies and memory resources vary significantly based on their hierarchical placement in the network. Coexisting in this environment is a dynamic set of mobile user devices, which generate computationally intensive and delay sensitive tasks. Each user device is heterogeneous in terms of its local computational capability and available battery energy. Furthermore, each device is equipped with multiple communication interfaces, enabling connectivity to different access points utilizing diverse radio access technologies. The system operates in a time slotted manner, where resource allocation decisions are computed and executed at the beginning of each discrete time slot. We assume a quasi static channel model, implying that the wireless channel conditions and the network topology remain relatively constant for the duration of a single time slot but may vary independently from one slot to the next.

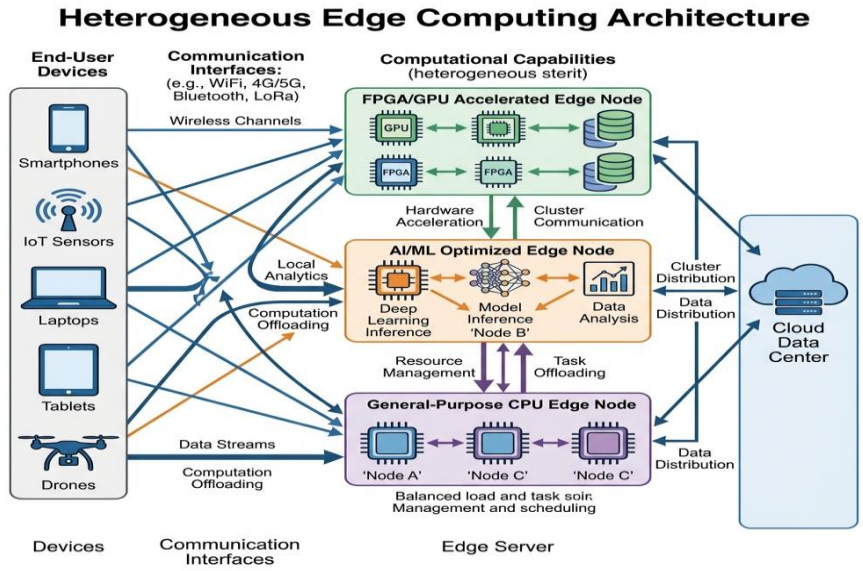



*Figure 1Comprehensive Schematic of the Heterogeneous Edge Computing Architecture*

*Table 1Primary Simulation Parameters and Architectural Configurations*

| Parameter Description | Mathematical Notation | Assumed Nominal Value |
|---|---|---|
| System Bandwidth Limit | B | 20 Megahertz |
| Background Noise Power Spectral Density | N_0 | -174 Decibel-milliwatts per Hertz |
| Maximum Device Transmit Power | P_max | 24 Decibel-milliwatts |
| Edge Server Computational Capacity | F_max | 50 Gigahertz |
| Local Device Computational Capacity | f_local | 2.5 Gigahertz |
| Task Input Data Size Range | D_in | 500 to 1500 Kilobytes |
| Processing Density Requirement | C_task | 1000 Cycles per Byte |

### 3.2 Communication Model and Interference Analysis

When a mobile device decides to offload its computational task to an edge server, it must transmit the task's input data over a selected wireless communication interface. The communication environment is subject to path loss, shadowing, and multi path fading. We define a binary decision variable that indicates whether a specific device selects a particular interface connected to a specific edge server. Because multiple devices may attempt to offload tasks simultaneously using the same radio access technology, the system must account for co channel interference. The signal to interference plus noise ratio for a transmitting device is determined by its allocated transmit power, the channel gain between the device and the selected access point, the aggregate interference from other transmitting devices utilizing the same frequency band, and the background additive white Gaussian noise. According to the Shannon Hartley theorem, the achievable uplink transmission data rate for a device is a logarithmic function of the signal to interference plus noise ratio and the allocated bandwidth. The transmission delay is subsequently calculated by dividing the total task input data size by the achievable data rate. Simultaneously, the energy consumed by the device during the transmission phase is the product of the

transmission power and the transmission delay. It is critical to note that increasing the transmission power reduces the transmission delay but exponentially increases the energy consumption and exacerbates the interference experienced by other devices in the network. Therefore, the communication resource allocation must carefully balance transmit power and bandwidth assignment to optimize the network wide performance metrics without violating individual power constraints.

**3.3 Computation Task and Delay Modeling**

The computational tasks generated by the user devices are modeled using two primary parameters: the size of the input data that must be processed and the processing density, which represents the number of central processing unit cycles required to process one bit of input data. The total computational workload of a task is the product of these two parameters. We adopt a partial offloading model, wherein a task can be arbitrarily divided into two independent subtasks. One subtask is executed locally on the mobile device's central processing unit, while the remaining subtask is offloaded to the edge server. The local execution delay is determined by the size of the local subtask and the device's local computational frequency. The local energy consumption is modeled as a function of the switched capacitance of the local processor architecture, multiplied by the square of the local frequency and the total number of required cycles. For the offloaded portion of the task, the execution delay at the edge server depends on the computational capacity allocated to that specific task by the server's hypervisor. Because edge servers handle requests from multiple devices simultaneously, the total allocated computational frequency across all hosted tasks cannot exceed the server's physical maximum capacity. The total delay experienced by the offloaded subtask encompasses the communication transmission delay, the queuing delay at the server, and the actual execution delay. Since the local and offloaded subtasks are executed in parallel, the total task completion time is determined by the maximum of the local execution delay and the edge execution delay. This parallel execution model necessitates precise coordination between the task splitting ratio, the communication rate, and the edge computational frequency allocation to prevent one component from becoming a severe bottleneck.

**3.4 Interface Selection and Energy Consumption Model**

The selection of a communication interface is inextricably linked to both the communication and computation models. Each interface possesses distinct characteristics regarding energy efficiency, latency overhead, and maximum allowable transmission power. For instance, selecting a cellular interface might offer broad coverage and high reliability but incur significant energy consumption due to the distance to the macro base station. Conversely, selecting a short range wireless local area network interface might require lower transmission power but offer limited bandwidth and higher susceptibility to local interference. The interface selection decision is mathematically represented by a multi dimensional binary matrix, subject to strict exclusivity constraints ensuring that each offloaded subtask utilizes exactly one interface per time slot to avoid unnecessary overhead and packet duplication. To capture the holistic performance of the system, we define a comprehensive utility function for each device, formulated as a weighted sum of the total task completion delay and the total energy consumption. The weighting parameters are customized for each device to reflect its specific operational priorities. A device with a critically low battery will assign a higher weight to energy conservation, whereas a device running a real time autonomous navigation application will heavily prioritize latency reduction. The global system objective is to minimize the aggregate utility function across all active devices in the network. We express the foundational mathematical optimization framework through the following constrained minimization problem:

$$min_{p,f,s} \sum_{i=1}^{M} (\omega_i^{t} T_i(p,f,s) + \omega_i^{e} E_i(p,f,s))$$

In this formulation, the variables represent the continuous allocation of transmit powers, the continuous allocation of computational frequencies at the edge servers, and the binary interface selection decisions, respectively. The functions calculate the total delay and total energy for each device, weighted by their respective preference parameters. This multi objective formulation encapsulates the extreme complexity of jointly orchestrating heterogeneous resources across diverse network components.

## 4. Joint Resource Allocation Methodology

### 4.1 Problem Formulation and Convexity Analysis

The optimization problem formulated in the previous section presents severe analytical and computational challenges. It is characterized as a mixed integer non linear programming problem due to the coexistence of continuous resource allocation variables, such as power and processor frequency, and binary integer variables representing the interface selection and task splitting decisions. Furthermore, the objective function is inherently non convex. The transmission delay component involves the inverse of the logarithmic Shannon capacity function, which is non convex with respect to the transmission power and bandwidth. The parallel execution model introduces a maximum function, which is continuous but non differentiable at the points where local and remote delays are equal. Additionally, the co channel interference terms in the denominator of the signal to noise ratio create strong coupling among the decision variables of different users, meaning that one device's power allocation directly alters the feasible solution space for all other devices. Directly applying standard convex optimization techniques, such as interior point methods, to this problem is impossible and would inevitably result in convergence to poor local minima. Exhaustive search methodologies for the binary variables, such as branch and bound, require exponential time complexity, rendering them entirely unsuitable for the strict real time constraints of dynamic edge computing environments. Therefore, a sophisticated mathematical transformation is required. We first address the non differentiability of the maximum function by introducing auxiliary variables and transforming the min max objective into a set of equivalent linear inequality constraints. This standard epigraph transformation linearizes the objective function at the expense of expanding the constraint space. Next, we address the highly coupled interference terms by utilizing the properties of fractional programming. Specifically, we apply the quadratic transform technique to decouple the numerator and denominator of the signal to interference plus noise ratio, effectively isolating the power variables and transforming the transmission rate expression into a more tractable, albeit still complex, mathematical structure.

### 4.2 Decomposition and Algorithmic Framework

To systematically solve the transformed non convex problem, we propose an iterative optimization framework based on the principles of block coordinate descent. The fundamental philosophy of block coordinate descent is to partition the complex set of optimization variables into smaller, independent blocks. The algorithm then iteratively optimizes the objective function with respect to one block of variables while holding all other blocks constant. This process is repeated sequentially across all blocks until the entire system converges to a stable solution. In our proposed methodology, we decompose the global problem into three distinct, manageable subproblems: computing resource allocation, communication power control, and interface selection. In the first phase of the iteration, we fix the transmission powers and interface selections, reducing the problem to the allocation of computational frequencies at the edge servers and local devices. Because the communication delay is fixed, this subproblem is rigorously proven to be strictly convex and can be efficiently solved using the method of Lagrange multipliers or standard convex solvers. In the second phase, we fix the computational allocations and interface decisions to optimize the transmission powers. Although the quadratic

transform decoupled the interference, this subproblem requires successive convex approximation to handle the residual non convexities in the rate function. By utilizing a first order Taylor series expansion, we construct a global lower bound surrogate function that is iteratively maximized. Finally, in the third phase, with power and computation variables fixed, we tackle the combinatorial interface selection problem. To avoid exponential complexity, we relax the binary integer constraints into continuous probability variables bounded between zero and one. We then incorporate a carefully designed penalty term into the objective function to force the relaxed continuous variables toward extreme binary values as the iterations progress.

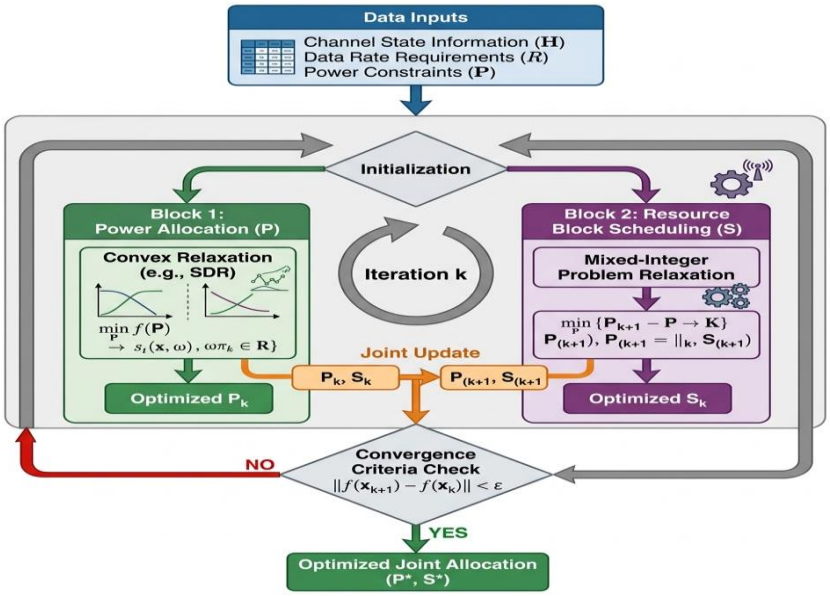



*Figure 2Algorithmic Flowchart and Resource Orchestration Sequence*

### 4.3 Distributed Optimization Implementation

While the centralized block coordinate descent algorithm provides an excellent theoretical foundation, requiring a single central controller to gather all channel state information, device parameters, and task requirements introduces significant signaling overhead and potential single points of failure. To enhance the scalability and robustness of our proposed methodology, we extend the algorithmic framework into a distributed optimization implementation utilizing the Alternating Direction Method of Multipliers. This technique allows the massive optimization problem to be fragmented and solved collaboratively across multiple edge servers and end user devices. We introduce localized consensus variables at each edge server and construct an augmented Lagrangian function. The function incorporates both linear penalty terms and quadratic penalty terms to enforce consistency between the local decisions made by individual devices and the global constraints managed by the edge infrastructure. The mathematical structure of this augmented Lagrangian is expressed as follows:

$$L(x,\lambda) = F(x) + \sum_{k=1}^{K} \lambda_k H_k(x) + \frac{\rho}{2} \sum_{k=1}^{K} (H_k(x))^2$$

In this formulation, the primary objective function is augmented with the Lagrangian multipliers and the squared constraint violation terms weighted by the penalty parameter. In the distributed execution, each mobile device independently optimizes its own transmission power and local task splitting ratio based on broadcasted shadow prices. Simultaneously, each edge server independently optimizes its internal computational frequency allocation. The updated variables are then exchanged, and the shadow prices are updated using dual subgradient ascent. This

distributed approach dramatically reduces the communication overhead required for optimization and naturally preserves the privacy of individual device parameters.

### 4.4 Computational Complexity and Convergence Analysis

The practical viability of any resource allocation algorithm in edge networks heavily depends on its computational complexity and convergence properties. We rigorously analyze the convergence of the proposed iterative block coordinate descent framework. Because the objective function is bounded from below by zero, and because each step of the block coordinate descent is mathematically guaranteed to either decrease or maintain the objective value, the sequence of objective values produced by the algorithm is monotonically decreasing and strictly bounded. According to the monotone convergence theorem, the algorithm is theoretically guaranteed to converge to a stationary point. Furthermore, the use of successive convex approximation in the power allocation subproblem satisfies the strict conditions required for convergence to a Karush Kuhn Tucker point of the original non convex problem. From a complexity standpoint, the centralized block coordinate descent algorithm requires a polynomial number of iterations to reach a predefined error tolerance. Within each iteration, the computing subproblem requires sorting operations, and the power allocation subproblem involves solving a constrained convex quadratic program. Table 2 provides a detailed comparative analysis of the theoretical complexity. Compared to exhaustive search methods which scale exponentially, and standard interior point methods applied to the monolithic problem which require cubic complexity with respect to the total number of variables, our decoupled approach reduces the complexity to a highly manageable quasi linear scale. This profound reduction in execution time ensures that the resource allocation decisions can be computed within the strict coherence time of the wireless channels, validating the framework for highly dynamic, real time implementations.

*Table 2Theoretical Computational Complexity of Various Allocation Schemes*

| Algorithm Type | Optimization Strategy | Big O Notation Complexity |
|---|---|---|
| Exhaustive Search | Combinatorial Branch and Bound | O(2^N * M^3) |
| Centralized Convex Relaxation | Interior Point Method on Monolithic Problem | O(N^3 * log(1/epsilon)) |
| Proposed Method | Iterative Block Coordinate Descent | O(I_max * N * log N) |

## 5. Experimental Results and Conclusion

### 5.1 Simulation Environment Setup

To rigorously evaluate the performance of the proposed joint computing, communication, and interface resource allocation methodology, we developed a comprehensive discrete event simulation environment using specialized mathematical modeling software. The simulated network topology consists of a heterogeneous deployment spanning a defined geographical area, comprising three distinct macro base stations acting as primary edge computing nodes, supplemented by six small cell access points providing localized, high bandwidth coverage. We deploy a variable population of mobile devices, ranging from twenty to one hundred users, randomly distributed across the coverage area following a spatial Poisson point process. The wireless channel models incorporate standard urban macrocell path loss paradigms, log normal shadowing with an eight decibel standard deviation, and Rayleigh fading for multipath effects. The computational tasks generated by the devices are modeled with input data sizes uniformly distributed between five hundred and fifteen hundred kilobytes, requiring processing densities varying from five hundred to one thousand cycles per byte. To establish a robust comparative

baseline, we benchmark our proposed methodology against three conventional resource allocation schemes prevalent in existing literature: a computation only optimization scheme that uniformly distributes communication resources, a communication only scheme that blindly allocates edge processing power, and a greedy heuristic interface selection algorithm that minimizes immediate path loss without considering holistic system capacity.

**5.2 Performance Evaluation on Delay and Energy**

The first set of simulation results focuses on the system wide aggregate utility, which represents the weighted sum of task completion delay and energy consumption. As the number of active mobile devices in the network increases from twenty to one hundred, the aggregate utility inevitably rises across all evaluated schemes due to intensified co channel interference and heightened contention for limited edge server computational cycles. However, the proposed joint allocation framework demonstrates a significantly slower rate of degradation compared to the baseline methods. At a network load of eighty devices, our proposed method reduces the aggregate system cost by approximately thirty five percent compared to the computation only optimization scheme, and by nearly forty five percent compared to the greedy heuristic approach. This substantial performance gap is directly attributable to the holistic nature of our algorithm. When network congestion peaks, the proposed methodology intelligently shifts the task splitting ratios, directing devices to execute more computations locally while simultaneously reallocating transmission powers to suppress severe interference bottlenecks. Conversely, the isolated optimization schemes quickly reach a state of catastrophic resource starvation, where devices transmit data rapidly only to have it queued indefinitely at the edge server, resulting in massive delay penalties and wasted transmission energy. We further isolate the specific metrics of average task delay and average device energy consumption. The simulation data clearly indicates that the proposed algorithm excels at balancing these conflicting objectives. Devices with stringent latency requirements successfully secured higher computational frequencies at the edge servers and optimized their transmit powers to ensure rapid data offloading. Meanwhile, devices configured to prioritize battery conservation effectively utilized local execution for smaller tasks and selected the most energy efficient communication interfaces, avoiding the heavy power drain associated with transmitting massive data payloads over degraded cellular links. The synergistic orchestration of multi dimensional resources explicitly prevents the formation of isolated bottlenecks that plague conventional, single domain management strategies.

**5.3 Scalability and Convergence Benchmarks**

To validate the theoretical complexity analysis presented in Section 4, we empirically measured the execution time and convergence behavior of the proposed algorithm under varying network densities. The iterative block coordinate descent framework demonstrated exceptional convergence speed, typically requiring fewer than fifteen iterations to reach a stable objective value within a strict error tolerance of ten to the power of minus four, regardless of the initial starting conditions. This rapid convergence is crucial for practical implementations where the network state changes dynamically. Furthermore, we evaluated the performance of the distributed Alternating Direction Method of Multipliers implementation. The results confirm that the distributed approach achieves an objective value within two percent of the centralized optimal solution while simultaneously reducing the control signaling overhead by an order of magnitude. The distributed framework scales almost linearly with the number of devices, confirming its applicability for massive internet of things deployments. The minor performance degradation observed in the distributed algorithm is a highly acceptable trade off for the massive gains in architectural scalability, fault tolerance, and preservation of user data privacy.

### 5.4 Concluding Remarks and Future Trajectories

In this paper, we have comprehensively investigated the critical problem of resource orchestration in highly heterogeneous edge computing environments. By recognizing the fundamental interdependencies between physical hardware, wireless channels, and network protocols, we formulated a rigorous joint optimization problem encompassing computing allocation, communication power control, and dynamic interface selection. To overcome the profound mathematical intractability of the resulting mixed integer non linear programming model, we developed an innovative iterative algorithmic framework leveraging block coordinate descent and successive convex approximation. Extensive theoretical analysis and simulation experiments have conclusively demonstrated that our proposed holistic methodology vastly outperforms traditional isolated resource management schemes, providing significant reductions in system wide delay and energy consumption while ensuring robust scalability. The results validate the necessity of breaking down traditional operational silos in network management to achieve true system optimization. Future research trajectories will focus on integrating proactive machine learning techniques, such as deep reinforcement learning, into the optimization framework to handle highly non stationary environments with unknown channel dynamics. Additionally, extending the system model to incorporate the mobility patterns of vehicular edge computing and the integration of low earth orbit satellite interfaces will be critical steps toward realizing the vision of ubiquitous, global edge intelligence.

### References


Chen, A., Li, X., Li, Y., Di, X., & Liu, X. (2021). A high-precision interface for tunneling magnetoresistance sensors with 0.15 nT/√ Hz resolution.Modern Physics Letters B,35(36), 2150527.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. Journal of Field Robotics, 42(5), 2318-2336.

Yang Z, Ji W, Guo Q, et al. CLAP: Cross-Layer Adaptive Pipelining Inference Scheduling for Resource-Efficient Edge-Cloud Vision Systems[J]. ACM Transactions on Architecture and Code Optimization, 2026, 23(2): 1-25.

Yao, Y., Zhu, Z., Miao, P., Cheng, X., Shu, F., & Wang, J. (2025). Optimizing hybrid RIS-aided ISAC systems in V2X networks: A deep reinforcement learning method for anti-eavesdropping techniques. IEEE Transactions on Vehicular Technology, 74 (6), 9224-9239.

Li, X., Hu, J., & Liu, X. (2020). Harmonic Distortion Optimization for Sigma-Delta Modulators Interface Circuit of TMR Sensors.Sensors,20(4), 1041.

Zhang, J., & Ding, J. (2026). CA-PPO-AXI: Coverage-Aware Deep Reinforcement Learning for Adaptive AXI Bus Protocol Verification.Fundamental Scientific Reports in Multidisciplinary Areas,2(02), 159-169.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. In Proceedings of SOSP.

Zhang, W., Li, M., & Chen, X. (2026). Distributed Edge-Fog-Cloud Architecture for Intelligent Fault Diagnosis of Permanent Magnet Synchronous Motors Using Temporal Fusion Transformers and Federated Learning. Fundamental Scientific Reports in Multidisciplinary Areas, 2 (02), 121-131.

Yang, W., Qin, Y., & Yang, Z. (2022). A reinforcement learning based data storage and traffic management in information-centric data center networks. Mobile Networks and Applications, 27(1), 266-275.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Li, X., Yin, L., Chen, W., Gao, Z., & Liu, X. (2017). A high-resolution tunneling magneto-resistance sensor interface circuit.Modern Physics Letters B,31(04), 1750030.

Xiao, W., Tan, Q., Chen, M., Yao, Y., & Shu, F. (2026). Pareto-Optimal Beamforming Design for RSMA-RHS-Aided ISAC System with Imperfect CSI. IEEE Transactions on Vehicular Technology.

Ou, J., Guo, J., Jiang, S., Li, X., Xue, R., Tian, W., & Buyya, R. (2025). Accelerating long-context inference of large language models via dynamic attention load balancing. Knowledge-Based Systems, 115018.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Yao, Y., Lu, Z., Chen, G., Huang, C., Feng, C., & Quek, T. Q. (2026). Interference Management in ISAC-SAGINs Based on Transformer-Enabled Mean-Field Reinforcement Learning Method. IEEE Transactions on Wireless Communications.

Xia, L., Yang, M., Wang, J., Yan, Z., Ren, Y., Yu, G., & Lei, K. (2025, September). Mamba4Net: Distilled Hybrid Mamba Large Language Models For Networking. In2025 IEEE 33rd International Conference on Network Protocols (ICNP)(pp. 1-10). IEEE.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.