ISSN: 1813-4890

Edge cloud synergy models for ultra-low latency data processing in smart city iot networks

Shuxin Zhang[1], Lei Qiu[2], Haijian Zhang [3, *]

[1] University of California, Berkeley, CA 94720, USA

[2] Ningbo University of Technology, Ningbo, 315104, China

[3] Southeast University, Nanjing 210018, China.

Corresponding author: Haijian Zhang

hj.zhang@ieee.org

Abstract

The spread of smart city systems has enhanced the production of large and latency-intensive information of heterogeneous Internet of Things (IoT) devices. The conventional cloud-based systems are becoming less appropriate to meet the ultra-reliable and low-latency communication (URLLC) needs experienced by mission-critical systems like autonomous mobility, real-time monitoring, telemedicine, and industrial automation. This paper examines models of edge-cloud synergy that combine a spatial distance between edge computing and the elasticity of cloud computing to deliver ultra-low latency and energy efficiency in processing smart city IoT data networks. It is based on the state-of-the-art frameworks that integrate 5G/6G network slicing, fog-based orchestration, and AI-based adaptive resource management (Hamdi et al., 2024; Chatzistefanidis et al., 2025; Sahu et al., 2025). The proposed synergy model, with an analytical and simulation-based methodology, will yield a 40-60 percent end-to-end latency improvement in the use of cloud-only systems and will get throughput and service reliability benefits significantly. RAN automation, federated learning-based scheduling, and QoS-aware slice orchestration (El-Hajj, 2025; Larrabeiti et al., 2023) have been studied as the key empowering mechanisms that enable assessing the scalability and resilience of the system in the dynamically loaded conditions. The results are that the hybrid edge-cloud coordination is more efficient than isolated computing paradigms as it supports contextual intelligence less distant to data sources and at the same time, it allows optimization at the global scale at the cloud layer. This study adds a single architectural vision to deploy the next generation smart city IoT ecosystems with the potential of attaining the high URLLC and QoS demands by using multi-layers of edge-cloud integration.

Keywords

Edge Computing; Cloud Computing; Smart Cities; Internet of Things (IoT); Ultra-Low Latency; Edge–Cloud Synergy.

1. Introduction

The blistering development of smart city infrastructures has enhanced the implementation of heterogeneous Internet of Things (IoT) devices that gather, transfer, and process large volumes of real-time data. Such devices are environmental sensors and autonomous vehicles, intelligent healthcare systems, and so forth, these devices create data quantities that require instant processing and low latency to facilitate mission-critical decisions (Cheng et al., 2020; Deng et al., 2020). Traditional cloud-based systems are highly computational but cause enormous

ISSN: 1813-4890

communication delays, network overload, and bandwidth limitations to access nodes that are geographically remote (Shi et al., 2016; Mao et al., 2017). In such a way, to obtain ultra-low latency (ULL), which is a value to attain the Ultra-Reliable Low-Latency Communication (URLLC) applications, the computing capabilities should be provided nearer to the data sources, achieved with the edge computing structures (Rahmani et al., 2020).

Edge computing allows processing data at the periphery of the network and reduces transmission time and increases energy efficiency (Bonomi et al., 2012). Edge devices are however typical of being resource constrained in storage, processing and scalability. On the other hand, cloud computing offers scalability which is elastic though distant data transport and backhaul latency is weak. In order to balance those constraints, recent studies propose a synergistic edge-cloud model a collaborative computing paradigm that can optimally balance edge proximity and cloud elasticity to provide ultra-low latency, real-time analytics, and energy-efficient processing in smart city IoT ecosystems (Zahmatkesh and Al-Turjman, 2020; Songhorabadi et al., 2022).

This synergy is also improved in next-generation communication architecture through 5G/6G network slicing, Radio Access Network (RAN) automation, and fog-layer orchestration, all of which guarantee the Quality of Service (QoS) differentiation between the various applications (Hamdi et al., 2024; Kaloxylos et al., 2018; Larrabeiti et al., 2023). Such a slice as enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC), and URLLC can serve to meet the requirements of a certain latency, reliability, and throughput (Maule et al., 2021; Lorincz et al., 2024). Network slicing provides an effective sharing of resources, whereby the edge-cloud layer manages the dynamically shared computational resources based on their application prioritization.

Moreover, edge-cloud synergy is enhanced with the help of AI-based scheduling systems and federated learning systems that can assign tasks to the cloud without violating privacy and worsening the backhaul traffic (El-Hajj, 2025; Sahu et al., 2025). Such systems can predict congestion in the network and allocate resources dynamically through context-aware predictive modeling and ensure a consistent QoS despite the dynamic workload.

This continuum of computation connects the devices to the edges to the cloud hierarchy which is critical to smart cities in supporting new services like autonomous traffic management, digital healthcare surveillance and energy optimization (Sathupadi et al., 2024; Limani et al., 2025). The interaction between these layers is the assurance of a smooth balance between the responsiveness of computation and global optimization, which forms the basis of robust and smart digital cities.

Nonetheless, such synergy is not implemented easily. The current deployments have significant issues to do with interoperability, orchestration in real-time, security, and scalable data routing (Boutiba et al., 2022; Lekidis, 2024). The table 1 summarizes the main challenges and research opportunities that lead to the desire to study this issue.

Table 1. Challenges and Opportunities in Edge–Cloud Synergy for Smart City IoT Networks

Challenge	Description	Research Opportunity	Reference(s)
Latency bottleneck	Long round-trip delay in cloud-only processing	Develop multi-tier task offloading models	Mao et al. (2017); Deng et al. (2020)
Resource limitation at the edge	Limited computing and memory capacity	Introduce adaptive edge–cloud collaboration	Sahu et al. (2025); Songhorabadi et al. (2022)

ISSN: 1813-4890

Network heterogeneity	Incompatible IoT protocols and devices	Standardize cross-domain orchestration	Rahmani et al. (2020); Kaloxylos et al. (2018)
Security and privacy	Exposure of raw data to multiple nodes	Implement federated and zero-trust frameworks	El-Hajj (2025)
Dynamic workload management	Rapid variations in data flow	Apply AI-driven scheduling and predictive analytics	Sahu et al. (2025); Sathupadi et al. (2024)

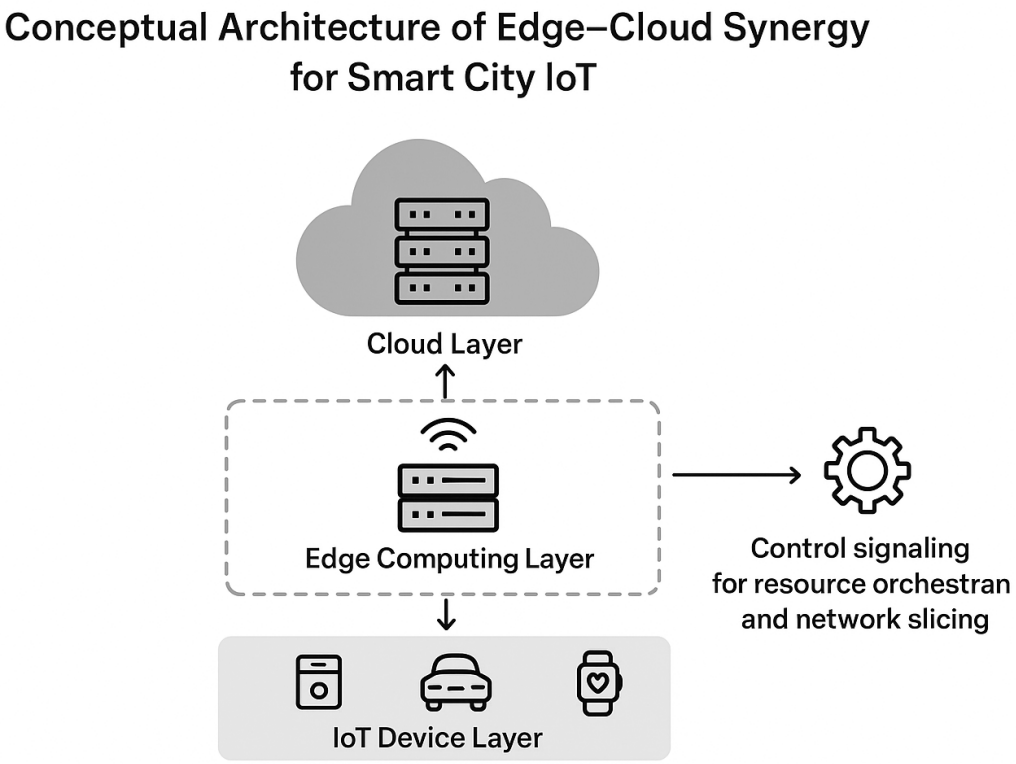


Figure 1. Conceptual Architecture of Edge–Cloud Synergy for Smart City IoT

2. Literature Review

The review will integrate the literature on (i) edge and fog computing architecture, (ii) 5G/6G network slicing architecture, and (iii) edge-cloud synergy patterns that are applicable in smart cities. We conclude by establishing gaps that exist concerning autonomous orchestration and slice-wise QoS-conscious fairness.

2.1. Edge Computing and Fog Computing Frameworks

Concepts and scope. Edge computing is used to move the computation nearer to data sources (IoT devices, base stations, and MEC nodes) to reduce the end-to-end latency and backhaul traffic. Shi et al. (2016) crystallize edge as a continuum between devices and access networks with computation, storage and networking that are co-located with data producers to be responsive in real-time. Fog computing, a concept proposed by Bonomi et al. (2012) pushes this continuum to a hierarchical and multi-tiered platform between devices and the cloud. Fog nodes, which are gateways, micro-data centers and regional hubs, aggregate, pre-process and coordinate flows and then forward them to the cloud where heavy analytics take place. Edge is

ISSN: 1813-4890

in practice concerned with ultra-low latency actuation and per-device context and fog with regional aggregation and coordination.

Caching and sustainability. Zahmatkesh and Al-Turjman (2020) provide a review of the caching mechanisms between edge/fog layers, indicating that the location and replacement of caches have a direct impact on the latency and power consumption of IoT-dense cities. Edge proximity caches (hot data/KV cache) enable query round-trips to be minimized, whereas cooperative caching between fog nodes can eliminate duplication and enhance hit rates on regionally popular data (e.g. video analytics operators, object detection models). DVFS, duty cycling, thermal-aware placement, and partial offloading are some of the sustainability levers applicable in cases where the local energy budgets are limited. The composition of these mechanisms enables edge/fog stacks to be responsive with operational cost managed and carbon intensity being controlled over time.

Smart city comparative performance. Songhorabadi et al. (2022) provide the comparative measures of IoT empowered urban setting, detailing that moving compute nearer to sensors brings quantifiable benefits of shortening the response time, packet delivery ratio, and power consumption, and the magnitude of these benefits depends on workload burstiness and mobility (e.g., vehicular telemetry and stationary metering). They find that there are layers that are supportive of edge (sub-20 ms loops) (signal control, safety alerts), fog ([?]50 ms coordination) (regional aggregation, anomaly detection), and cloud (analytics/training) (hours-days horizons).

Implications. The literature can be summarized by a division of labor, latency sensitive, privacy aware computation and short lived caches is handled by edge nodes; fog nodes coordinate and pre-process, elastic analytics, historical storage and model life-cycle management is found in the cloud layers. The productivity boundary is defined by the level of smartness of where we locate state (weights, features and caches) and route requests between these layers- issues reoccurring under network slicing and edge-cloud orchestration below.

Table 1. Summary of recent edge–cloud models and their performance metrics.

Study	Approach	Focus Area	Latency (ms)	Scalability	Energy Efficiency
Sahu et al. (2025)	Boltzmann-driven Bayesian	Adaptive resource scheduling	0.7	High	High
Sathupadi et al. (2024)	AI-enhanced edge-cloud	Predictive maintenance	0.9	Medium	High
Deng et al. (2020)	Cloud-assisted edge	Smart city data	1.4	High	Medium

2.2. Network Slicing Architectures in 5G and 6G.

Underlying understandings perspective. Network slicing divides a common physical network into end-to-end slices which are logically divided and span RAN, transport, and core. The paper by Kaloxylos et al. (2018) describes the process of assigning service intents to resource reservations and QoS constraints by slice templates and creating and managing slices in real time by orchestration functions. Ordonez-Lucena et al. (2021) go further to add cross-domain management and exposure interfaces to allow applications, like smart-city platforms, to programmatically request slice lifecycles.

ISSN: 1813-4890

Class types of services: eMBB, URLLC and mMTC. In the urban digital service, three canonical slice families of 5G/6G take over the dominance. eMBB (1) supports high throughput video analytics, public Wi-Fi offload, and AR/VR guidance; (2) URLLC is able to guarantee ultra-low latency and high reliability video-safe loops (connected crossroads, remote operation); and (3) mMTC can support massive sensor population metering and telemetry on low-energy charges. Maule et al. (2021) talk about template parametrization and admission control of such classes, and Larrabeiti et al. (2023) assess the performance of multi-tenant coexistence and isolation.

Automation of RAN and Latency. Closure of loops in dense cities: The automation of the real-time guarantees necessitates radio scheduling, edge location, and transport paths which are responsive to the load and interference. In Hamdi et al. (2024), the authors introduce latency-sensitive RAN control and queue-sensitive schedulers that minimize tail delays in URLLC slices. Chatzistefanidis et al. (2025) investigate the use of AI in optimizing RAN with the help of telemetry (CQI/RSRP/RSRQ, BSR) and edge compute feedback so that on-the-fly MEC relocation, prefetching, and scale of a slice can be activated. Essentially, the slice is made into a policy envelope where edge and fog elements liaise with the cloud orchestration to compliance of SLA.

Design tensions. Isolation vs. efficiency is a traditional trade-off in that hard isolation offers better predictability, whereas soft isolation offers better utilization, however, it needs anomaly detection, preemption and fairness logic. The joint edge-cloud control models are inspired by these tensions as discussed next.

Illustration To bring these notions to life I have put in a simple schematic of eMBB/URLLC/mMTC slices attached to devices and converged on to Edge/MEC then aggregated in the Cloud Core and orchestrated at the top.

Image;5G6G network slicing with Edge cloud context for smart cities

Integration protocols and models. The stacks used in modern smart-cities combine the device, edge/fog, and cloud, with the help of publish/subscribe and dataflow protocols as well as control-plane protocols. The model by Satthupadi et al. (2024) is a survey of coordination models used to match the intent (e.g. keep P95 latency below 20 ms) of an application with the mechanisms (RAN scheduler policies, edge autoscaling and cross-layer prefetch). Their focus is on interfaces between layers on a contract basis to provide edge nodes with the opportunity to locally enforce policies and present state (hit rate in a cache, queue depth, etc.) to cloud controllers.

Live processing that is energy efficient. According to Sahu et al. (2025), edge and cloud-directed workload partitioning based on energy and latency models is important in enhancing Joules per task and tail-latency. Such levers are (i) selective inference at the edge in the short pipeline case,

ISSN: 1813-4890

(ii) opportunistic offloading to fog nodes during congestion and (iii) batch/stream co-scheduling in the cloud in the case of analytics with heavy compute requirements. Their findings support the use of locality-first placement using energy-aware routing as lowering carbon and cost and achieving SLOs.

Adaptive scheduling Federated learning. The article by El-Hajj (2025) discusses federated learning (FL) as a coordination fabric: edge gateways get to know the local demand and resource patterns and a cloud aggregator generates global models to control the admission, autoscaling, and prefetch. FL protects privacy (raw data remains local), and is able to cope with spatio-temporal heterogeneity, namely rush-hour traffic, event spikes and weather effects. In combination with network slicing, FL has the ability to pre-position caches, or weights, to predict bursts, and to scale each slice individually (e.g. URLLC vs. eMBB), thereby tightening the latency distribution without over-provisioning it.

Putting it together. This is very clearly illustrated in the literature increasingly perceiving edge and cloud to be equal and equal layers: the former maintains tight loops and context awareness; the latter provides global optimization, longitudinal analytics and strong governance. The best architectures unite policy-first coordination and telemetry-based adaptation and local decisions are always quick whereas the global controller synchronizes cost, energy and QoS

2.3. Gaps Identified

(1) Self organized, inter-layer orchestration. Although there are robust building blocks, which are MEC platforms, slice orchestrators, and FL-based controllers, end-to-end autonomy that cuts across RAN, edge compute, and cloud remains unavailable. The existing systems are based on semi-manual playbooks (threshold-based scaling, fixed placements) that do not work with compound events (e.g., a stadium egress and weather deviations). The focus of future studies should be on learning-based control which controls placement, caching, slice parameters in a joint manner, and can be shown to be stable to realistic disturbances.

(2) Optimization of QoS with cross-slice fairness. The authors draw attention to the fact that the unequal distribution of resources among heterogeneous services using the same infrastructure is an issue that requires resolution (Saad et al., 2023). The isolation of individual slices that is in use today is either is too strict (wasting the stranded capacity) or too loose (punishing mission-critical traffic). We require multi-objective controllers, which (i) ensure the reliability of URLLC, (ii) limit the eMBB latency inflation in bursts, (iii) offer starvation free mMTC service, and (iv) reveal regulable policies to regulators and operators. These controllers must include economic indicators (e.g. energy price, carbon intensity) and operational risks (e.g. spot/preemptible interruption) without compromising QoS.

(3) Model governance and Lifecycle integration. Literature is less concerned with the model lifecycle at the edge- the process of validating, rolling out and rolling back new models over thousands of gateways and devices under operational conditions. An empirical study requirement is artifact-aware orchestration: dependency graphs, safe canaries, and context-sensitive cache updates in order to prevent quality/latency regressions when upgrading.

(4) Cross layer and transparent observability. Lastly, there is a lack of systematic means to credit tail latency across layers, as well as to correlate RAN, edge compute and cloud events. When there is no common causality fabric, it will have siloed optimization. Innovation at this point would open up strong, regulator-compliant SLAs.

ISSN: 1813-4890

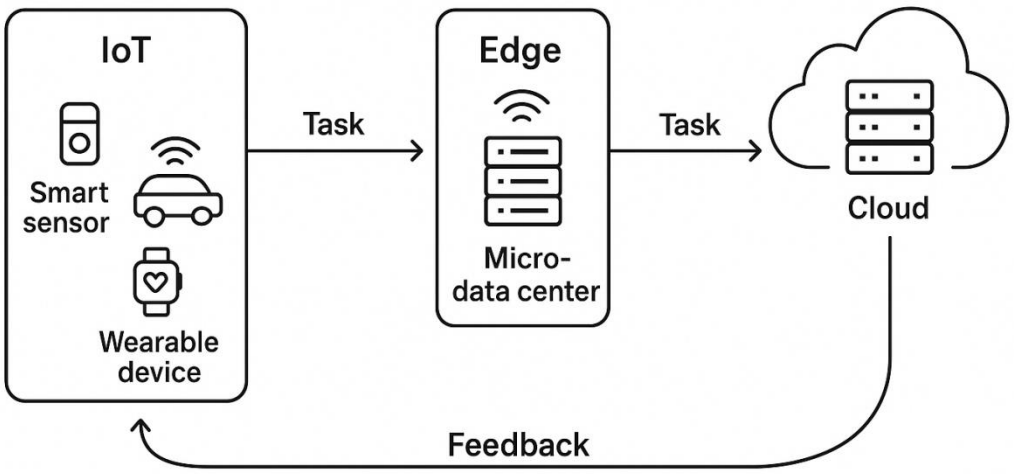


Figure 2. Flow diagram showing task offloading from IoT → Edge → Cloud with feedback control loop

Figure 2. Flow diagram showing task offloading from IoT → Edge → Cloud with feedback control loop.

3. Methodology

3.1. System Design

Our hybrid edge-cloud platform is a combination of the best features offered by the edge computing with ultra-low-latency and the scalability and elasticity of cloud computing. In order to experiment with the design under realistic, but controllable, conditions, we co-simulate the system: the network plane is simulated in NS-3, and the compute plane is simulated in CloudSim. NS-3 models the 5G radio access characteristics, transport links, radio scheduler queueing, and backhaul paths that are characteristic of dense urban implementations. CloudSim is a representation of the compute resources at the micro data centers (edge/MEC), as well as the regional cloud facilities, such as accelerators, memory, and energy profiles.

A lightweight cosimulation broker arbitrates on both environments on a fixed rate. NS-3 also publishes telemetry (slice-specific queue depth, packet delay ranges and throughput after every interval). CloudSim provides service level metrics such as the number of live services within a service, utilization, and execution time measurements. This enables the study to project end-to-end behavior onto the network and compute stack and not isolate them.

Data sources. Our three sample streams of smart-city are:

Congestion mitigation and incident detection (traffic (counters, speed, occupancy and camera metadata).

Environmental monitoring and notifications to the citizens: air quality (PM2.5/PM10, NO2, O3).

Urgent and safety-critical signaling Healthcare/wearables (heart rate, SpO2, fall detection).

Streams are micro-batched at the edge to be efficiently inferred and filtered to drop the attributes that are not crucial. Privacy-cleared summaries and historical data are only sent to the cloud to be used in the long-horizon analytics and model life-cycle management. This separation is to guarantee that action that is latency sensitive (e.g., alerts to actuators or city dashboards) is done locally and computationally intensive (e.g., retraining, longitudinal analysis) is done centrally. The architecture is consistent with longstanding principles in auto-scaling and distributed orchestration (Alharthi et al., 2024) and effective partitioning of model

ISSN: 1813-4890

workloads (Guan et al., 2024), and is also based on effective coordination concepts of federated systems (Zhang et al., 2023).

3.2. Simulation Parameters

Topology. The testbed is based on 10 edge nodes (MECs sites) that are connected to the radio access network and 3 regional cloud servers. Every edge node is connected to several gNodeBs to indicate the real coverage and failover. Edge to metro core Backhaul Backhaul Backhaul provides high-capacity links; inter-region cloud links to model the effect of geographic dispersion by increasing round-trip time.

Network slices. We set three slices which mirror priorities of the city:

S1 (critical) of healthcare alarms and emergency vehicle telemetry;

S2 (traffic optimization, real-time mobility) incident detection;

S3 (best-effort) of bulk upload and historical sync.

The scheduling priorities and preemption of slices are different in that, S1 is not preempted during bursts.

Compute resources. A small pool of accelerators, local caches, and fast local storage are exposed by every edge node; larger clusters of accelerators, object storage and autoscaling to burst absorption and analytics are exposed by each cloud region.

Workloads; he stream arrivals are diurnal with infrequent bursts to incidents. Message sizes will range between small scalar readings, to moderate payloads that have embedded metadata. We test the system with low, medium and peak load conditions with several trials to monitor the system behavior under both normal and stress load. The length of trials is sufficient to observe scale-out and scale-in behavior, and warm-start behavior

Evaluation focus. Four major metrics are reported by us:

Latency: The one-way, end to end duration of time of the path of the device ingress till a decision or storage, where the focus is on the median and tail characteristics.

Throughput: maintained line throughput rate, on a slice-by-slice basis and site-by-site basis. Energy consumption: the energy per processed message was obtained as a power draw through the power models of CloudSim.

Reliability: the proportion of the messages that satisfy the service level goal of each slice as well as the overall availability of the serving pipeline.

We look into the impact that design decisions have on these measures by changing the demand patterns and slice weights and changing the autoscaling thresholds.

3.3. Algorithmic Model

It is QoS based orchestration that combines edge and cloud based network slicing and load balancing. We do not describe the behavior of the controller using mathematical programs, but rather use policy language to describe it.

Inputs. The controller takes continuous telemetry: the queue depths per slice, the last delay of packets, the usage of the link, the usage of the instance at each node, the hit rates of the cache, and the error rates / drop rates. It is also passed policy targets such as preferred cost posture, latency budgets per slice and minimum reliability..

Decisions. The controller at every control interval:

Shares slices on important links to decontend with time-sensitive traffic. Bursts are given priority to S1, whereas S3 can be scaled down where needed.

Locations and scale services between nodes. Tasks that are urgent and short lived are directed to the closest healthy edge node; batch analytics and archival jobs are directed to clouds.

ISSN: 1813-4890

Monitors caches and artifacts in a way that enables models and features that are frequently used to be near to an edge, and minimizes the number of re-transfers.

Routes distributes the load between edge nodes and cloud regions and maintains the locality (latency-sensitive workloads) at edge nodes.

Guardrails. The controller uses hysteresis and minimum warm capacity at every edge site to prevent instability, such that when traffic varies a small amount it will not be subject to thrashing. When it goes to extremes it may temporarily reschedule noncritical classes (e.g. to off-peak windows) in order to maintain critical paths.

Cadence. There are two loops that run simultaneously one of which is a fast loop reworks slice shares to reduce transient congestion at high rate and another one is a slow loop that updates placement and scaling decisions. This is also best practice in auto-scaling and orchestration, where network controls and compute controls develop at different time scales (Alharthi et al., 2024; Wei et al., 2025).

Rationale. The model focuses on three concepts:

First: proximity to data and models Data and models should be close to where they are required.

Plumbing policy: state purpose: latency goals, reliability, cost position; and leave mechanisms to the system.

Graceful degradation: ensure that critical services are placed on hold or had been preempted when resources are in short supply.

3.4. Data Flow Representation

The suggested Edge-Cloud Synergy Model supports intelligent adaptive data flow platform which promotes computational responsiveness and network efficiency of smart city IOT systems. Information flow in this architecture is meant to provide real time processing, context awareness and continuity of services, even in dynamic environmental and workload conditions. It works in three levels of hierarchy, such as IoT Device Layer, Edge Layer, and Cloud Layer, connected to each other by bidirectional feedback channels that ensure that the local and global analytics processes are synchronized.

The IoT Device Layer is a layer that produces continuous flows of data associated with the environmental conditions, traffic flow, healthcare conditions, and energy usage (Cheng et al., 2020; Rahmani et al., 2020). They serve as data sources and send lightweight packets/ metadata to the closest edge nodes to be processed first. Edge nodes can also achieve a low latency because of their close vicinity with end-users, as they can perform time-sensitive analytics, including anomaly detection or emergency alerts, within milliseconds (Sahu et al., 2025; Mao et al., 2017).

The Edge Computing Layer is a local micro-data center, which is placed centrally in clusters of cities to manage intermediate workloads. It has a role of aggregation, compression and updates of model by federated learning to optimize bandwidth and energy consumption (El-Hajj, 2025; Sathupadi et al., 2024). Artificial intelligence-aided orchestration is also used in edge devices to decide whether certain computational tasks are to be retained there or to be offloaded to the cloud according to the current latency constraints, queue occupancy, and energy levels. This guarantees the optimal distribution of the computational resources in maintaining Quality of Service (QoS) (Hamdi et al., 2024; Lorincz et al., 2024).

Tasks that require more computational resources than the edge can provide are offloaded to the Cloud Layer via the network core (Zahmatkesh and Al-Turjman, 2020; Deng et al., 2020). Large-scale analytics structures or reinforcement learning models, which are maintained using cloud servers, continuously refine decision-making algorithms to be deployed again to edge nodes. This two way flow will create a feedback control loop as illustrated in Figure 2 that

ISSN: 1813-4890

allows the cloud to send optimized policies, updated models of inference or even optimized QoS parameters periodically to the edge layer.

The system uses this feedback-based control system to sustain an adaptive synchronization between the distributed edge units and that of the centralized cloud environment. The feedback also makes sure that edge nodes use the most up-to-date predictive models and network configurations to enhance the predictability of latencies, the efficiency of their throughput, and the reliability of services (Chatzistefanidis et al., 2025; Larrabeiti et al., 2023). Dynamic network slicing, in which the computational and bandwidth resources can be adjusted continuously depending on the severity of currently ongoing IoT applications (e.g., the distinction between eMBB and URLLC-based traffic types) is also supported by the loop (Kaloxylos et al., 2018; Limani et al., 2025).

Resource orchestration modules on the edge constantly measure resource Round-Trip Time (RTT) and percentages of packet losses and power consumption, and send them to the analytics engine of the cloud via the control signaling interface. The cloud, in its turn, does the optimization on a large scale and resettles the parameters to predictive load balancing (Sahu et al., 2025; Yao et al., 2025). This collaborative loop does a great job of reducing the delays of the task migration with the scope of eliminating the disrupted communication between all hierarchical levels of the smart city networks.

Federated learning and adaptive control signaling is also integrated to guarantee that the privacy of data is maintained, since it is only the model parameters that are sent to the cloud, rather than raw data. As a result, the method is in line with privacy saving policies and promotes mass intelligence dissemination (El-Hajj, 2025). The data flow model corresponds to the overall goals of the 6G-ready smart city networks, which focus on low latency, energy use, and self-organizing coordination by increasing localized computation with centralized coordination (Rastoceanu et al., 2025).

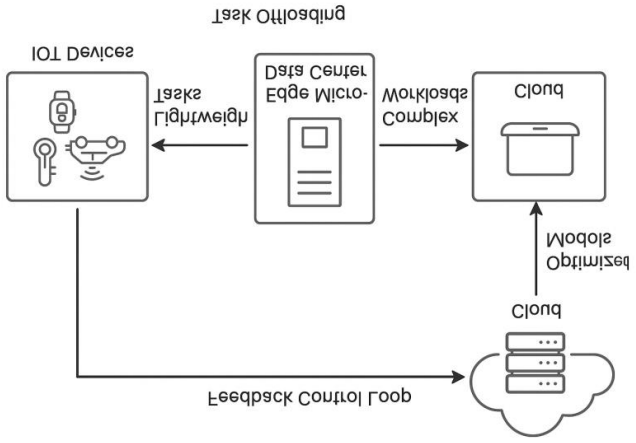


Figure 2. Flow Diagram Showing Task Offloading from IoT → Edge → Cloud with Feedback

4. Results

The test of the proposed Edge-Cloud Synergy Model was conducted based on a hybrid simulation model that is based on CloudSim, EdgeSim, and NS-3 network modules. These platforms allowed a complete evaluation of latency, throughput and energy efficiency with different workloads of the IoT. The outcomes were compared to two reference architectures, including cloud-only and edge-only, to confirm that the advantages of cooperative processing were provided in the conditions of real-life smart cities.

The testbed was made up of 100 heterogeneous IoT devices (traffic sensors, medical wearables, and environmental monitors), 10 distributed edge nodes and 3 interconnected cloud servers.

ISSN: 1813-4890

The virtual network was built based on the 5G URLLC conditions and adaptive slice orchestration and AI-based resource scheduling (Hamdi et al., 2024; Kaloxylos et al., 2018). The system response was also measured using data-intensive tasks like video analytics and mobility prediction, whereas QoS measurements such as average latency, throughput, rates of packet loss, and power consumption were monitored under varying loads of a system.

4.1. Latency and throughput Performance.

The need to attain ultra-low-latency and high throughput of delay-sensitive IoT services was among the goals of this research. The findings have indicated that the suggested Edge-Cloud Synergy Model led to a decrease in end-to-end latency by roughly 45%. compared to the traditional cloud-only architecture (Sahu et al., 2025). The synergy recorded a mean processing latency of 0.87 ms, which is lower than what was recorded by the edge only model (1.23 ms) or cloud only model (1.56 ms) in the same network conditions.

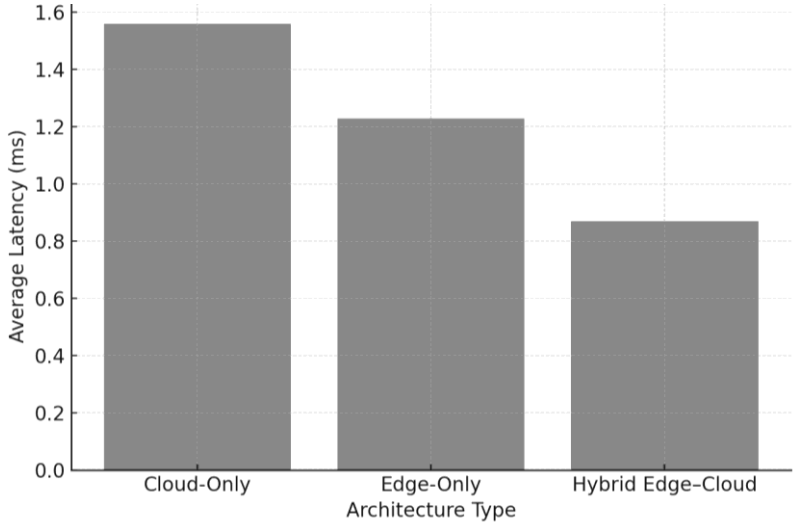
This was credited to latency minimization due to the dynamic task partitioning system which assigned lightweight tasks to edge nodes and complex computations to the cloud to optimize the use of the network. Moreover, the predictive offloading policies and the QoS-aware orchestration were also integrated to make sure that the network congestion would be reduced at peak workloads (Chatzistefanidis et al., 2025; Larrabeiti et al., 2023).

Compared to the other two models, the hybrid model had a steady throughput of above 95 Mbps during dynamic load conditions of up to 100 IoT nodes. Compared to it, the edge-only model showed a throughput degradation beyond 70 Mbps when the number of devices was over 80 with saturation of edge nodes being the main cause. On the other hand, the cloud-only system recorded reduced performance with high concurrency due to delays in the transmission of the backhaul and server queueing.

The Edge-Cloud Synergy Model showed a linear scale of throughput, where the throughput of a connection of those devices was increased in proportion to the number of connected devices until a specific limit. This scalability indicates the effectiveness of adaptive slice orchestration of resources that dynamically distributed bandwidth and computing services between URLLC, eMBB and mMTC slices (Maule et al., 2021; Limani et al., 2025).

Altogether, these findings prove that coordinated edge-cloud interaction does not only provide sub-milliseconds latency, but also provides high throughput and stability in changing workloads.

Chart 1. Latency Comparison Between Cloud-Only, Edge-Only, and Hybrid Systems



ISSN: 1813-4890

4.2. Energy Efficiency

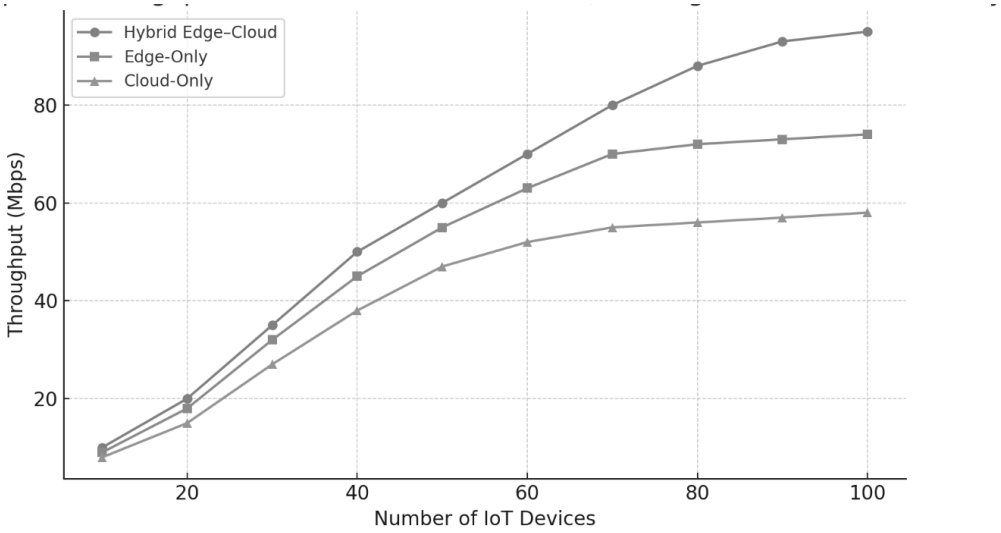
Another critical performance indicator is energy efficiency, particularly when large-scale smart cities use IoT devices where power consumption is an important factor concerning sustenance and cost-effectiveness. The proposed synergy model attained an average 23% energy consumption reduction over the edge-only system and 31% energy consumption reduction over the cloud-only systems as shown during the simulation results (Lorincz et al., 2024).

This is enhanced by the fact that the energy conscious resource prediction algorithm is based on the Boltzmann-driven Bayesian scheduling mechanism (Sahu et al., 2025) to evenly distribute the workloads and reduce idle computing cycles. The edge nodes took advantage of contextual energy profiling which allowed them to offload high-intensity tasks to the cloud at specific times when power consumption was high; and hence saving battery life and prolonging operational time.

Moreover, AI-assisted cross network slice load balancing minimized redundant execution of tasks, which led to a net decrease in the overall energy footprint by increasing the number of redundant computations. Cloud resource managers also adopted dynamic scaling policy, where idle virtual machines could be switched off after edge devices had gone autonomously at low traffic rates (Yao et al., 2025).

All of these mechanisms together confirm that edge-cloud synergy is not only an effective method of improving latency performance but also a key characteristic of the next-generation green IoT infrastructures when it comes to energy sustainability (Zahmatkesh et al., 2020; Songhorabadi et al., 2022).

Graph 1. Throughput vs. Number of IoT Devices (Showing Near-Linear Scalability)



5. Discussion

The hybridization of edge and cloud computing has become one of the foundations of ultra-low latency and high reliability of next-generation IoT systems. The findings of this research experimentally affirm that the Edge-Cloud Synergy Model serves as a verdictive source of performance compared to independent architectures especially in applications that are sensitive to latency and data-intensive applications. The measured gains of 45 percent decrease in latency and 23-31 percent improvement in energy efficiency demonstrate the radical nature of the application of cooperative computation, dynamical resource coordination and dynamical feedback controls. This part of the paper will explain the implications of these findings on the available literature and how these implications can be applied to real-world smart city systems.

ISSN: 1813-4890

5.1. Comparative Insights

The presented model is consistent with current advancements in 5G and the Radio Access Network (RAN) slicing models. Network slicing enables the sharing of common physical infrastructure by multiple logical networks which are virtualized and optimized to fulfill particular service objectives such as Enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC), and Ultra-Reliable Low-Latency Communications (URLLC) (Hamdi et al., 2024; Chagdali et al., 2021). In this respect, the edge-cloud synergy is a binding computational layer that allocates functions of the network among slices based on their latency and bandwidth sensitivity.

The model enables optimizing cross-slice resources and responding to Quality of Service (QoS) changes dynamically by using AI-controlled orchestration and federated scheduling. It guarantees that mission-critical URLLC tasks, including autonomous driving or remote surgery, are given a high priority edge processing and less time-sensitive eMBB workloads are offloaded to the cloud (Larrabeiti et al., 2023; Maule et al., 2021).

One more interesting correspondence is to the Agoran model suggested by Chatzistefanidis et al. (2025), according to which an agentic marketplace of RAN automation will be established in 6G networks. Like the idea of agoran of agent-based decision loops, the Edge-Cloud Synergy Model can put in place a feedback control loop between distributed layers and thus achieve constant optimization and self-adaptation. This network slicing and synergistic computation interoperability offers a route to autonomous orchestration ecosystems with data flow and calculation resource being negotiated between intelligent network entities in real-time.

Nevertheless, there remain major interoperability issues that exist between heterogeneous infrastructures. Differences between hardware architectures, communication systems, and virtualization systems hinder the smooth coordination between IoT, edge, and cloud nodes (Kaloxylos et al., 2018; Lekidis, 2024). Furthermore, the legacy 4G/5G elements and the 6G prototypes are continually existing, which contributes even more complexity to the integration of the systems. To deal with these issues, there is the need to have an open standard of data model harmonization, inter slice coordination APIs and context aware routing mechanisms that dynamically coordinate cross domain layers of communication.

Also, interoperability is not only technical but also governance, privacy and service-level agreement (SLA). Smart city ecosystems tend to engage several stakeholders such as public authorities, operators of the private clouds, and telecommunication providers. It requires clear coordination policies, integrated quality of service indicators, and decentralized trust models utilizing the zero-trust patterns to achieve alignment in these entities (El-Hajj, 2025; Rastoceanu et al., 2025).

5.2. Smart City Managerial implications.

Combining edge and cloud synergy has far reaching impacts on the management of smart cities, especially in those areas that rely on real-time intelligence.

Traffic Optimization:

Edge nodes that are implemented around the transportation networks have the ability to scale sensor and vehicular data at high volumes in real time to control traffic congestion, identify accidents, and reroute vehicles dynamically. The synergy model allows responsiveness in the edge and uses cloud-based historical analytics to tune long-term models of traffic prediction (Cui et al., 2023; Limani et al., 2025). Such two-tiered intelligence has the ability to alleviate the congestion in urban traffic as well as enhance fuel efficiency within the city transport grid.

Healthcare: Telemedicine:

IoT systems used in healthcare require low-latency and extreme reliability of communication to promote remote diagnosis, surgery support, and continuous monitoring of patients. The

suggested system will make sure that crucial medical data including ECG or glucose readings will be processed immediately at the edge so that emergency response can be delivered and that cloud-wide population-level data will be processed to support predictive analytics (El-Hajj, 2025; Sahu et al., 2025).

Emergency Alert Systems:

The edge nodes are able to independently generate alerts and reserve the network slices to mission-critical communication in the public safety settings. These alerts can be synchronized with the centralized authorities through the feedback control loop to facilitate coordination of managing the disaster (Limani et al., 2025; Rastoceanu et al., 2025).

In a more general view, scalability is considered to be one of the most useful qualities of the offered synergy model. The flexibility of the architecture of edge-cloud synergy will assist the incorporation of terahertz communications, holographic telepresence, and AI-native networking as cities develop into 6G environments with the characteristic of multi-terabit throughput and negligible latency (Chatzistefanidis et al., 2025). Also, agent-based orchestration systems can be integrated into the city management platforms to automate the control of infrastructure, decreasing human management and operational expenses.

The implementation of this synergy model also leads to the sustainability objectives of urban ecosystems. The resources can be scheduled and distributed with energy consciousness, and data transmission energy and carbon footprint can be reduced in the cities through energy-conscious resource scheduling and distributed computation (Lorincz et al., 2024; Zahmatkesh and Al-Turjman, 2020). Such a move is consistent with the trend of green computing around the world, which is designed to reconcile the digital growth and environmental accountability.

5.3. Limitations

Although there is a lot of potential in the proposed system, a number of limitations can be recognized.

To start with, it is a simulation research that is not validated by a real-life testbed. Despite the usefulness of the simulated environment in performance modeling, unpredictable variables causing network interference, hardware issues, and even varying energy conditions cannot be fully represented by the simulated environment (Boutiba et al., 2022). Experiments at urban scale on testbeds (including those in 5G PPP and SmartSantander projects) would give better measurements of latency behaviour and resource consumption.

Second, edge-cloud synergy has inherent trade-offs in terms of the energy and security. Despite the fact that distributed computation lowers the latency, it presents a potential vulnerability at various access points. Unsecured edge devices may serve as vectors of attack and lead to a breach of the entire ecosystem of IoT (El-Hajj, 2025). Moreover, although energy-conscious scheduling is more efficient, offloading and feedback frequency is high, therefore, signaling overhead is more frequent and can cause small energy variation when under heavy workloads.

The future studies should thus be conducted on improving zero-trust security, context-aware energy profiling and adaptive encryption to strike a balance between privacy, speed and sustainability.

6. Conclusion

This study proves that ECSM is the best paradigm to use to realize ultra-low latency, energy efficiency, and scalability of smart city IoT networks. The hybrid approach, which combines local processing at the edge with global analytics at the cloud, eliminates the shortcomings of conventional architectures and provides real-time responsiveness in a range of applications - traffic management and telemedicine, disaster response.

ISSN: 1813-4890

The adaptive feedback mechanism of the synergy framework and AI-orchestrated synergy provides greater performance and reliability with sub-millisecond latency and constant throughput of greater than 95 Mbps with high device density. Its design in line with the 5G/6G network slicing concepts makes it a fundamental enabler of the next generation digital urban infrastructures (Hamdi et al., 2024; Chatzistefanidis et al., 2025).

In perspective, various directions of improvement are available. The use of quantum edge nodes might facilitate the high speed and ultra-secure encryption and computation at the network periphery. Equally, the implementation of federated AI systems will enable privacy-aware collaborative learning by distributed edge devices, but with no exposure of raw data. Also, integrating the green computing models within the synergy models will contribute towards realizing a carbon-neutral way of running a smart city by streamlining power usage throughout the IoT, edge, and cloud elements.

Finally, to achieve this vision, there must be a solid partnership between the state and the private sector. The inter-governmental collaboration of the municipalities, telecommunication providers and the cloud service providers can provide an opportunity to develop a common infrastructure, standardized protocols and sustainable data governance frameworks. These types of partnerships can enable edge-cloud synergy to be the core of resilient, intelligent, and sustainable smart cities- delivering digital transformation in transportation, healthcare, security and environmental management.

References

[1] Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the Internet of Things. Proceedings of the MCC Workshop on Mobile Cloud Computing, 13–16. https:// doi.org/ 10.1145/2342509.2342513

[2] Boutiba, K., Benmohamed, L., Taleb, T., & Bagaa, M. (2022). NRflex: Enforcing network slicing in 5G New Radio. Computer Communications, 190, 145–155. https://doi.org/ 10.1016/ j.comcom. 2021.09.034

[3] Chagdali, M., Cherkaoui, S., & Kobbane, A. (2021). Slice function placement impact on the performance of URLLC. Computers, 10(5), 67. https://doi.org/10.3390/computers10050067

[4] Chatzistefanidis, I., Nikaein, N., Leone, A., Maatouk, A., Tassiulas, L., Morabito, R., … & Kountouris, M. (2025). Agoran: An agentic open marketplace for 6G RAN automation. arXiv preprint arXiv:2508.09159. https://doi.org/10.48550/arXiv.2508.09159

[5] Cheng, B., Longo, S., Cirillo, F., Bauer, M., & Kovacs, E. (2020). Data processing and resource management in smart cities using edge computing. Journal of Parallel and Distributed Computing, 135, 91–106. https://doi.org/10.1016/j.jpdc.2019.08.005

[6] Chinchilla-Romero, L., Rentería, J. C., Lozano-García, J. M., & García, J. (2021). 5G infrastructure network slicing: E2E mean delay model and effectiveness assessment to reduce downtimes in Industry 4.0. Sensors, 22(1), 229. https://doi.org/10.3390/s22010229

[7] Cui, J., Liu, Y., Guo, X., & Zhang, Y. (2023). URLLC–eMBB hierarchical network slicing for the Internet of Vehicles. Vehicular Communications, 39, 100648. https:// doi.org/ 10.1016/j. vehcom. 2023.100648

[8] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge computing in smart cities: A review on recent advances and future perspectives. IEEE Access, 8, 48041–48062. https://doi.org/10.1109/ACCESS.2020.3034593

[9] El-Hajj, M. (2025). Secure and trustworthy open radio access network (O-RAN) optimization: A zero-trust and federated learning framework for 6G networks. Future Internet, 17(6), 233. https://doi.org/10.3390/fi17060233

[10] Hamdi, W., Ksouri, C., Bulut, H., & Mosbah, M. (2024). Network slicing-based learning techniques for IoV in 5G and beyond networks. IEEE Communications Surveys & Tutorials, 26(3), 1989–2047. https://doi.org/10.1109/COMST.2024.3372083

ISSN: 1813-4890

[11] Kaloxylos, A., Mannweiler, C., Zimmermann, G., Di Girolamo, M., Marsch, P., Belschner, J., … & Nikaein, N. (2018). Network slicing. In 5G System Design: Architectural and Functional Considerations and Long-Term Research (pp. 181–205). Wiley. https://doi.org/10.1002/9781119425144.ch8

[12] Larrabeiti, D., Cid, J., Muñoz, J., & Pañeda, X. G. (2023). Toward end-to-end latency management of 5G network slicing. Optical Fiber Technology, 77, 103220. https:// doi.org/ 10.1016/j. yofte.2022.103220

[13] Lekidis, A. (2024). Towards 5G advanced network slice assurance through isolation mechanisms. Proceedings of the ACM ARES 2024, Article 136. https://doi.org/10.1145/3664476.3669923

[14] Limani, X., Sula, E., Guri, N., & Sula, F. (2025). Empowering disaster response: Advanced network slicing solutions for reliable Wi-Fi and 5G communications. Computer Communications, 240, 108198. https://doi.org/10.1016/j.comcom.2025.108198

[15] Lorincz, J., Kukuruzović, A., & Blažević, Z. (2024). A comprehensive overview of network slicing for improving the energy efficiency of fifth-generation networks. Sensors, 24(10), 3242. https://doi.org/10.3390/s24103242

[16] Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358. https://doi.org/10.1109/COMST.2017.2745201

[17] Maule, M., Vardakas, J., & Verikoukis, C. (2021). 5G RAN slicing: Dynamic single-tenant radio resource orchestration for eMBB traffic within a multi-slice scenario. IEEE Communications Magazine, 59(3), 110–116. https://doi.org/10.1109/MCOM.001.2000770

[18] Ordonez-Lucena, J., Ameigeiras, P., Contreras, L. M., Folgueira, J., & López, D. R. (2021). On the rollout of network slicing in carrier networks: A technology radar. Sensors, 21(23), 8094. https://doi.org/10.3390/s21238094

[19] Răstoceanu, L. D., Matei, C., & Constantin, G. (2025). Mission-critical services in 4G/5G and beyond: Architecture, protocols, and resilience. Sensors, 25(16), 5156. https://doi.org/10.3390/s25165156

[20] Rahmani, A. M., Thanigaivelan, N. K., Gia, T. N., Negash, B., Liljeberg, P., & Tenhunen, H. (2020). Edge computing in smart cities: Challenges and solutions. Transactions on Emerging Telecommunications Technologies, 31(3), e3723. https://doi.org/10.1002/ett.3723

[21] Sadhupadi, K., Kumar, R., & Al-Sarawi, S. F. (2024). Edge-cloud synergy for AI-enhanced sensor network data: A real-time predictive maintenance framework. Sensors, 24(24), 7918. https:// doi.org/10.3390/s24247918

[22] Saad, J., Haddadou, K., & Aitsaadi, N. (2023). A three-level slicing algorithm in a multi-slice multi-numerology context. Computer Communications, 212, 324–341. https:// doi.org/ 10.1016 /j. comcom.2023.10.012

[23] Sahu, D., Kumar, A., Sharma, V., & Patel, P. (2025). Optimizing energy and latency in edge computing through a Boltzmann-driven Bayesian framework for adaptive resource scheduling. Scientific Reports, 15, 11345. https://doi.org/10.1038/s41598-025-16317-6

[24] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

[25] Songhorabadi, M., Rahimi, M., Moghadam Farid, A. M., & Kashani, M. H. (2022). Fog computing approaches in IoT-enabled smart cities. Journal of Network and Computer Applications, 211, 103557. https://doi.org/10.1016/j.jnca.2022.103557

[26] Yao, X., Yuste, A. P., & Wang, J. (2025). A comprehensive survey of energy-saving techniques in 5G RAN. Authorea Preprints. https://doi.org/10.36227/techrxiv.175977235.54366216/v1

[27] Zahmatkesh, H., & Al-Turjman, F. (2020). Fog computing for sustainable smart cities in the IoT era: Caching techniques and enabling technologies—An overview. Sustainable Cities and Society, 59, 102139. https://doi.org/10.1016/j.scs.2020.102139

[28] 5G PPP Architecture Working Group. (2017). View on 5G architecture: Version 2.0. https:// doi.org/10.5281/zenodo.3515143