Markov model of controlled load distribution in the edge-IoT subsystem*

Viacheslav Kovtun[1,†], Oleh Yasniy[2,*,†] and Oleh Kovaliuk[1,†]

[1] Vinnytsia National Technical University, Khmelnytske shose, 95, Vinnytsia, 21021, Ukraine
[2] Ternopil Ivan Puluj National Technical University, Ruska St, 56, Ternopil, 46001, Ukraine

Abstract

The article presents a hybrid Markov model of adaptive load distribution within edge-IoT subsystems operating under conditions of stochastic traffic and heterogeneous environments. The proposed approach formalises the set of admissible and critical system states and introduces a routing policy with a probabilistic guarantee of stabilisation within QoS-defined configurations. For the first time, an integral efficiency criterion is proposed, combining the probability of remaining in desirable states with the minimum probability of stabilisation under perturbations. The model enables the development of adaptive routing strategies that minimise servicing costs and enhance resilience against node degradation and traffic fluctuations. Experimental results demonstrate the superiority of the proposed approach over classical strategies (round-robin and random) in terms of average delay, resource utilisation, and the loss function. The model has practical relevance for 6G edge architectures, autonomous systems, eHealth, and critical IoT infrastructures.

Keywords

edge computing, IoT, load balancing, Markov process, adaptive routing, QoS guarantee, stochastic modelling, network resilience

1. Introduction

Between 2024 and 2025, there has been an exponential increase in the number of IoT devices connected to global telecommunication networks, particularly in the domains of logistics, transport, healthcare, manufacturing, and smart cities [1, 2]. According to Statista, the number of active IoT connections is expected to exceed 30 billion by the end of 2025, a substantial proportion of which will operate based on edge computing architectures. This trend is driven by the need to process data as close as possible to its source to minimise latency, conserve bandwidth, and reduce energy consumption. Under such conditions, stable and adaptive management of computational resources at edge nodes becomes a critical determinant of Quality of Service (QoS). Notably, real-world incidents in the field of intelligent transport systems highlight the vulnerability of edge-IoT subsystems to overload conditions [3]. For instance, during the CES-2024 festival in Las Vegas, the smart traffic light system (relying on edge-level data processing from motion sensors) temporarily lost stability due to an unforeseen surge in traffic, resulting in disruptions to urban infrastructure. Such cases underscore the critical importance of dynamic request redistribution among edge nodes, capable of adapting to sudden changes in both load and network topology. In military and mission-critical security scenarios, such as autonomous reconnaissance systems, edge-IoT subsystems must maintain QoS even in the event of partial network disconnection or node degradation. For this reason, leading research institutions, including IEEE, 3GPP, and ITU, identify adaptive load management at the edge level as one of the key challenges for future 6G networks. Nevertheless, most existing models remain insensitive to the stochastic nature of traffic or fail to account for the

*CITI'2025: 3rd International Workshop on Computer Information Technologies in Industry 4.0, June 11–12, 2025, Ternopil, Ukraine
1,* Corresponding author.
† These authors contributed equally.
kovtun_v_v@vntu.edu.ua (V. Kovtun); o_yasniy@tntu.edu.ua (O. Yasniy); oleh.kovalyuk@vntu.edu.ua (O. Kovaliuk)
0000-0002-7624-7072 (V. Kovtun); 0000-0002-9820-9093 (O. Yasniy); 0000-0002-0718-010X (O. Kovaliuk)

variable performance of nodes in heterogeneous environments. Consequently, against the backdrop of rapidly evolving telecommunication infrastructures, increasingly stringent QoS requirements, and energy constraints, the development of flexible and perturbation-resilient load management models for edge-IoT subsystems is of paramount importance from both theoretical and applied perspectives.

Existing research in the field of load management within fog and edge-IoT environments encompasses a wide range of approaches, which may be broadly categorised by methodology [4–6]: heuristic and metaheuristic algorithms, probabilistic logics and game-theoretic strategies, graph-based methods for load structuring, hybrid computational schemes, as well as stochastic models based on Markov processes. Despite the increasing volume of publications in this area, critical analysis reveals that none of these directions provides an adequate level of formalised controllability required for the stable operation of heterogeneous edge-IoT systems with guaranteed QoS levels.

Most implemented solutions are based on stochastic or metaheuristic schemes [7–9], including particle swarm algorithms, hill climbing, genetic algorithms, and iterative load-shedding methods. These approaches are primarily aimed at reducing average service time, minimising latency, or improving energy efficiency. They have shown favourable results in simulation environments and constrained scenarios; however, nearly all are grounded in simplified models and do not provide a formal description of the system's admissible state space or the probabilistic dynamics of transitions between states. As a result, they lack guarantees of stabilisation within acceptable boundaries, particularly under conditions of traffic variation or node degradation.

Nowadays, a lot of problems in various fields of science and technology can be solved by means of probabilistic methods [10, 11]. This is also true for the above-mentioned problem. Probabilistic and fuzzy logic approaches [12–15] (including fuzzy rule-based controllers, fuzzy load-balancing systems, and game-theoretic models utilising Shapley value-based resource allocation) enable the implementation of adaptive local policies. These methods provide flexible responses to local changes; however, they are predominantly focused on the micro-level and do not account for the global dynamics of the system. The absence of a description of admissible and critical configuration sets, along with the uncertainty of transition trajectories within the state space, limits their applicability in systems characterised by high reliability requirements.

Graph-based methods [16–18] (such as vertex-cut, dynamic partitioning, or hot data caching) demonstrate computational efficiency in distributed networks with many nodes. However, their implementation typically assumes a static graph structure and does not account for temporal dynamics, which limits responsiveness to peak loads or partial infrastructure failures. As a result, these methods do not ensure guaranteed system behaviour under real-time conditions.

Hybrid approaches [19–21] (combinations of neural networks, caching algorithms, adaptive routing, and heuristic planning) have demonstrated promising results in specific domains such as smart city applications or eHealth. However, most of these models remain complex for mathematical analysis, lack unified loss evaluation criteria, and are insufficiently interpretable for safety-critical applications. Moreover, they rarely support structural modelling of system dynamics at the level of the admissible state space.

A separate category should be allocated to studies employing Markov Decision Processes (MDP) [22–25], including partially observable models (POMDP), constrained Markov models (CMDP), and even continuous-time variants (CTMDP). Some of these models, for instance, describe nodes as two-state systems with the optimisation of offloading policies based on index evaluation. Others apply reinforcement learning built upon CMDP to enable energy-efficient control while maintaining QoS. Nevertheless, most such models are either too simplified to capture multidimensional queuing behaviour or too complex for practical implementation (particularly in cases where input data are unstable or limited). Consequently, they either fail to reflect the actual heterogeneity of the system or exhibit limited applicability due to the necessity of training on large-scale datasets.

A common issue across all the examined classes of models is the absence of a unified efficiency criterion that simultaneously accounts for temporal characteristics, probabilistic risks of transitioning into critical states, deviation from target configurations, and the load on key nodes. Existing optimisation strategies remain fragmented: one approach seeks to minimise delay, another focuses on energy consumption, while a third targets the reduction of redirected requests. The lack of an integrated metric of systemic efficiency precludes a comprehensive evaluation of control quality and significantly complicates the synthesis of adaptive strategies.

Another drawback of most existing approaches is that they are based on the assumption of infrastructure homogeneity. In many models, nodes are treated as identical in their characteristics, even though real-world edge-IoT networks exhibit significant heterogeneity (including mobility, unstable power supply, and variability in cloud access). Under such conditions, disregarding this non-uniformity results in inadequate routing decisions, increased risk of overload, or degradation of QoS in the weaker nodes of the system.

It is also worth emphasising that most of the described control systems are designed as offline policies, relying on a priori knowledge of traffic patterns or user behaviour. Such approaches are unsuitable for operation under real-time conditions, where continuous policy updates are required based on the observed states of queues and nodes. The inability to reactively adapt to environmental changes significantly undermines the system's reliability under load.

These shortcomings become particularly critical in domains where even short-term degradation in service quality may lead to irreversible consequences (notably in autonomous transport, defence sensor networks, eHealth infrastructure, or distributed energy systems). In such cases, approaches are required that not only minimise average losses but also ensure, with probabilistic confidence, that the system remains within admissible operational boundaries.

This comprehensive analysis highlights the necessity of developing models that integrate a structured stochastic foundation, adaptive routing, risk evaluation across a range of configurations, and the capability to operate in real time within a heterogeneous environment. A model of this kind should not only capture average service efficiency but also ensure controllable system behaviour by employing formalised loss criteria aligned with the target objectives of QoS assurance in critical IoT architectures.

The object of the study is the process of adaptive load distribution in an edge-IoT subsystem, considering the stochastic nature of traffic and the heterogeneity of resources.

The subject of the study is a Markov model with controllable routing policies, which describes the state dynamics of the edge-IoT subsystem under conditions of variable load, critical configurations, and adaptive response to perturbations.

The aim of the study is to develop a mathematically grounded stochastic model for adaptive control of request distribution in edge-IoT subsystems, ensuring system stabilisation within QoS-controlled configurations and the minimisation of integral costs.

To achieve this aim, the following research objectives were set and addressed:

– To formalise the space of admissible and critical states of the edge-IoT subsystem, considering node heterogeneity and QoS parameters;

– To construct an adaptive Markov model for load control based on transitions between configurations with varying levels of performance;

– To define integral metrics of control efficiency that combine the probability of remaining in admissible states with the risk of degradation;

– To develop algorithms for generating hybrid routing policies capable of ensuring guaranteed stability even under traffic fluctuations;

– To conduct simulation analysis and comparison with classical approaches (round-robin and random distribution), evaluating delay, resource utilisation, and losses.

The main contribution of this study lies in the development of a formalised Markov model for adaptive load control in edge-IoT subsystems, which integrates the system's stochastic dynamics with a hybrid service structure and enables the subsystem to be maintained within QoS-defined configurations. The proposed approach ensures not only the minimisation of service costs but also

a high probability of avoiding critical states, even under variable traffic conditions, structural heterogeneity, and partial resource degradation. Unlike classical models, it allows for the quantitative incorporation of risks associated with transitions to critical configurations, supports the modelling of phase dynamics of stabilisation, and enables the construction of adaptive routing strategies with a guaranteed level of efficiency, as confirmed by the results of simulation analysis.

2. Models and methods

2.1. Formalisation of the State Space and Criteria for Optimal Load Distribution in the Edge-IoT Subsystem

The focus of this study is the formalisation of a model for controlled load distribution within an edge-IoT subsystem. The subsystem consists of a set of edge computing nodes E_k, $k \in I = \{1,2,\ldots,N_e\}$, where $N_e \in \mathbb{N}$ denotes the total number of edge nodes, and I is the corresponding index set. Each edge node E_k is modelled as an M/M/1 queueing system [26], with a service rate $\lambda_k \in \mathbb{R}_+$, which represents the average number of requests processed by the node per unit of time.

The management of request flows between edge nodes is carried out according to a routing policy defined by a probabilistic transition matrix $\Pi = [\pi_{km}]$, $\sum_{m=1}^{N_e} \pi_{km} = 1 \quad \forall k \in I$. The matrix Π is a square matrix of dimension $N_e \times N_e$, where each element $\pi_{km} \in [0,1]$ represents the probability that a request from node E_k will be forwarded to node E_m, or processed locally (in the case of $k = m$).

The current state of the subsystem at a discrete time $t \in \mathbb{Z}_+$ is described by a queue vector $q^{(t)} = \left(q_1^{(t)}, q_2^{(t)}, \ldots, q_{N_e}^{(t)}\right)$, where $q_k^{(t)} \in Z_+$ denotes the number of requests awaiting processing at a given edge node E_k at time t, and the vector $\vec{q}^{(t)} \in \mathbb{Z}_+^{N_e}$ characterises the overall load configuration of the entire edge-IoT subsystem at time t. The set of all admissible subsystem states is denoted by $Q \subseteq \mathbb{Z}_+^{N_e}$. Its cardinality is indicated as $c_Q = |Q|$, and the set of indices corresponding to the admissible states is referred to as $B = \{1,2,\ldots,c_Q\}$.

To formalise the desired functional state of the subsystem, a target vector $q^o = \left(q_1^o, q_2^o, \ldots, q_{N_e}^o\right) \in Q$ is introduced, where $q_k^o \in \mathbb{Z}_+$ denotes the desired number of requests at edge node E_k, corresponding to an ideal (balanced) load distribution. The vector q^o may be defined in accordance with the selected objective, such as minimisation of delay, energy consumption, or uniformity of processing.

The quality of each admissible state $q^{(n)} \in Q$, $n \in B$, is evaluated using a cost function

$$\Psi(n) = \sum_{k=1}^{N_e} \alpha_k q_k^{(n)} \qquad (1)$$

.

The parameter $q_k^{(n)} \in \mathbb{Z}_+$ introduced in expression (1) represents the number of requests at node E_k in state $q^{(n)}$, while $\alpha_k \in \mathbb{R}_+$ is a weighting coefficient that accounts for node priority, processing

cost, or energy consumption constraints. The function $\Psi(n) \in \mathbb{R}_+$ serves as an aggregated metric of load or service cost for the entire subsystem in the state indexed by n.

The objective of the edge-IoT subsystem's operation is to reach a state in which the costs are minimised:

$$\min_{n \in B} \Psi(n) \tag{2}$$

.

In addition to optimising the average load, it is also essential to evaluate boundary scenarios related to system overload. For this purpose, a maximum cost function $\Psi(1) = \max_{1 \le n \le c_Q} \Psi(n)$ is introduced, where $\Psi(1)$ denotes the value of the cost function (1) for the configuration $q^{(n)}$ in which the value of $\Psi(n)$ is the highest among all possible configurations. This approach enables the assessment of the most critical state of the subsystem and is of key importance in the design of QoS threshold mechanisms or protective mechanisms of the edge controller.

To construct load control policies in an edge-IoT subsystem, it is necessary to define subsets of subsystem state configurations that satisfy specific criteria of proximity to the target load distribution. Let $Y \subseteq Q$ denote the set of states considered admissible in terms of QoS, that is, those in which the load on each edge node remains within a controlled deviation from the target value. Let $Z = Q \setminus Y$ represent the set of undesirable states, in which the subsystem operates with efficiency reduced relative to the optimum. The number of elements in the defined sets is denoted as $c_Y = |Y|$, $c_Z = |Z|$, respectively.

The proximity of each admissible state $q^{(n)} \in Q$ to the target vector q^o is evaluated using normalised tolerance thresholds. To this end, a tolerance vector $\vec{b} = (b_i) \in \mathbb{Z}_+^{N_e}$ is introduced, where each element $b_i > 0$ defines the maximum permissible deviation of the queue at edge node E_i from the corresponding target value q_i^o: $q^{(n)} \in Y \Leftrightarrow \left| q_i^{(n)} - q_i^o \right| \le b_i \quad \forall i \in I$. The deviation threshold $b_i \in \mathbb{Z}_+$ for node E_i may be interpreted as an overload allowance or buffer depth within QoS constraints. Accordingly, the set Y constitutes a Euclidean neighbourhood of the vector q^o with a radius defined by the vector b, and includes all admissible configurations in which the overload at no edge node exceeds the critical level.

Let M denote the set of deterministic actions within the defined control policy, where each action regulates the selection of request redirection probabilities between the nodes of the edge-IoT subsystem. For each state $q^{(n)} \in Q$, a routing strategy is selected in the form of a vector $p(n) = (p_{nm})_{m=1}^{c_Q}$, which belongs to the space of probability distributions $p(n) \in P = \left\{ p_{nm} \ge 0, \sum_{m=1}^{c_Q} p_{nm} = 1 \right\}$, where $p_{nm} \in [0,1]$ represents the probability of transitioning from state $q^{(n)}$ to state $q^{(m)}$ under the selected control action, and P is the standard simplex in c_Q dimensions.

Thus, a control strategy in the edge-IoT subsystem is defined as a mapping $p : Q \to P$, according to which each subsystem state is associated with a probability distribution over subsequent states. Let $P_Y \subseteq P$ denote the set of such strategies that guarantee the subsystem

remains within admissible states. This implies that, under $q^{(n)} \in Y$, all transitions must keep the subsystem within the set Y:

$$P_Y = \left\{ p \in P \middle| p_{nm} = 0 \forall q^{(n)} \in Y, q^{(m)} \in Z \right\}. \tag{3}$$

Formally, the task of synthesising a control policy consists in determining a mapping $p(\cdot) \in P_Y$ that ensures the reachability of the set Y from any initial state of the subsystem, guarantees that the subsequent evolution of the subsystem remains within the boundaries of Y, and either minimises the cost function $\Psi(n)$ on average or ensures its boundedness. This formulation enables the problem to be formalised within the framework of a Markov process with a variable policy.

2.2. Hybrid Markov Model of the Edge-IoT Subsystem with Adaptive Routing Policies

Within the formalised model of controlled load distribution, which describes the behaviour of the edge-IoT subsystem as a stochastic process over the set of admissible configurations, it is essential to examine the conditions for guaranteed stabilisation of the subsystem within the target set Y. Despite the existence of policies that restrict transition probabilities to undesirable states in accordance with condition (3), the Markovian nature of the dynamics necessitates additional justification of stability in the presence of perturbations or random overloads.

Let $\mathrm{J} = \{1, 2, \ldots, c_Z\}$ denote the set of configuration indices belonging to the domain $Z = Q \setminus Y$. For each $J \in \mathrm{J}$, the quantity $\xi_J \in [0,1]$ is considered, representing the probability of reaching configuration $q^{(c_Y + J)} \in Z$ from any admissible state, as well as the vector $\tilde{\mu}_J \in \mathbb{R}_+^{N_e}$, which reflects the guaranteed service intensity at the corresponding edge nodes in these states.

To quantitatively assess the stability of the edge-IoT subsystem, a value φ is introduced, representing the lower bound of the probabilistic guarantee of stabilisation. It is defined as the minimum among the values $\min(\xi_J, \tilde{\mu}_J)$ for all $J \in \mathrm{J}$:

$$\varphi = \min_{J=1,\ldots,c_Z} \left\{ \min(\xi_J, \tilde{\mu}_J) \right\}. \tag{4}$$

From a practical implementation perspective, the value φ serves as a parameter used as a lower bound in constructing a control strategy that not only minimises the cost function according to criterion (2) but also ensures that the subsystem remains within admissible boundaries even in the presence of traffic fluctuations or partial degradation of computational resources. Incorporating the parameter $\tilde{\mu}_J$ as a dynamic indicator of the processing capacity of edge nodes enables the model to reflect scenarios in which the subsystem must operate under conditions of priority-driven resource reallocation, loss of channel capacity, or reduced energy autonomy at certain nodes. It is the combination of stochastic information regarding the likelihood of reaching configurations from the set Z, together with guarantees of local service availability, that allows a transition from static approaches to the construction of adaptive policies aimed at ensuring the stable operation of the edge-IoT subsystem within the set Y.

The further development of the model requires consideration of the hybrid nature of the service environment, in which edge nodes operate with dynamically varying service intensities under

stochastic traffic distribution. For each node E_i, we consider an extended approximation of the guaranteed service intensity in a degraded state $q^{(c_Y+J)}$. In this context, the configuration of the edge-IoT subsystem in each state $n \in B$ is described by a vector $\vec{s}(n) = \left(s_i^{(n)}\right) \in \mathbb{Z}_+^{N_e}$, where $s_i^{(n)}$ denotes the number of queued requests at node E_i in state n.

Additionally, two auxiliary quantities are introduced: $h_i = s_i^{(c_Y+J)} + \mu_i \varphi$ and $n_i = s_i^{(c_Y+J)} + \upsilon_i \varphi$, where υ_i represents an adjusted characteristic of the dynamic processing reserve aligned with routing paths. To compute this value, the quantity $L \in \mathbb{N}$ is used, denoting the number of edge nodes in the subsystem, along with the routing coefficients $\theta_{ji} \in [0,1]$, which describe the probability that a request processed at node E_j will be redirected to node E_i. Accordingly, the value υ_i is calculated as $\upsilon_i = \sum_{j=1}^{L} v_j \theta_{ji}$, where v_j is the processing intensity at edge node E_j, identified with $\breve{\mu}_j^J$.

On the basis of h_i and g_i, we introduce a hybrid estimate of service intensity $\breve{\mu}_i^J$, which equals $\left(h_i - s_i^o\right)/\varphi$ if $h_i > s_i^o$ holds, and μ_i in all other cases. Similarly, the parameter $\breve{\mu}_i^J$ is defined, constructed on the basis of g_i. These parameters will subsequently be used to construct the generator of the hybrid Markov process $\breve{A}^J = \left(\breve{a}_{mn}^J\right)$, where for each pair of states $m \neq n$ corresponding to configurations $s^{(m)}$ and $s^{(n)}$, the matrix $\breve{A}^J$ element is given by $\breve{a}_{mn}^J = \varepsilon\left(s_i^{(m)}\right)\tilde{\mu}_i^J v_{ij}^{(m)}$, and the diagonal elements are defined as $\breve{a}_{nn}^J = \sum_{i=1}^{L} \varepsilon\left(s_i^{(m)}\right)\tilde{\mu}_i^J$. Here, the function $\varepsilon(\cdot)$ equals 1 if the corresponding component is greater than zero, and 0 otherwise.

The temporal dynamics of the edge-IoT subsystem operation are characterised by the transition probability matrix $\breve{P}^J(t) = \exp\left(\breve{A}^J t\right)$. For each J, an indicator vector $\breve{q}^J$ is constructed, in which a unit value corresponds only to the component $c_Y + J$. The probability of reaching this state by time t is given as $F_J(t) = \breve{q}^J \breve{P}^J(t) e_1$, where e_1 is the unit vector representing the initial state of the subsystem, and the hybrid cost function is defined by expression

$$\eta_J^* = \int_0^{\varphi} t dF_J(t) + \varphi\left[1 - F_J(\varphi)\right] \quad (5)$$

.

Finally, over the set Q, a stationary distribution π_n is considered, which satisfies the normalisation condition and the Kolmogorov equations using the matrix $\breve{P}^J(t)$.

The analysis of the behaviour of the controlled load distribution policy in the edge-IoT subsystem concludes with the examination of the probability of remaining within the admissible set of configurations Y, previously defined as those corresponding to an acceptable deviation from the target state. In the context of the stochastic model, this probability is appropriately considered an integral indicator of the effectiveness of the constructed routing policy. Let $\pi(Y)$ denote the total stationary probability that the subsystem resides in any state $q^{(n)} \in Y$. This value is computed

as the sum of stationary probabilities π_n over all indices $n \in B_Y \subseteq B$, where B_Y is the set of indices corresponding to configurations Y: $\pi(Y) = \sum_{n \in B_Y} \pi_n$. The value $\pi(Y) \in [0,1]$ may be interpreted as a load management quality indicator, as it reflects the proportion of time the edge-IoT subsystem spends, on average, in admissible states.

To enhance the adaptability of the policy under conditions of high traffic variability, a refined criterion is introduced, according to which a generalised efficiency function $\eta = \varphi\pi(Y)$ is defined. Here, φ, as described by formula (4), characterises the lower bound of the probabilistic guarantee of reaching an admissible state, while the product $\eta \in [0,1]$ is regarded as an integral controllability criterion of the edge-IoT subsystem, accounting for both its current structure and probabilistic risks. This value may serve as an objective function in optimisation tasks related to routing and the identification of overloaded segments within the subsystem.

In cases where the assumptions regarding input flow intensities do not conform to a Poisson distribution, the model can be generalised by incorporating empirically obtained state transition frequencies or by computing $\pi(Y)$ numerically based on simulation results. In this way, even within a heterogeneous environment characterised by irregular topology and dynamically varying resource availability, the Markov model presented in this section enables the formulation of a stable balancing policy with a minimal probability of QoS degradation.

Given the hybrid nature of the edge-IoT subsystem and the set of reduced-performance states Z, it is important not only to determine the integral efficiency metric (see η), but also to construct localised evaluations of the cost function, which enable the individualisation of routing strategies according to the criticality of each specific performance degradation scenario.

For each index $J \in \{1,\ldots,c_Z\}$ corresponding to a specific reduced-performance state $q^{(c_Y+J)}$, a value $\eta_J \in \mathbb{R}_+$ is defined, representing the expected losses of the subsystem associated with entering that state. This value generalises the cost function η_J^*, previously introduced in formula (5), by incorporating the local transition structure, recovery probabilities, and the residual time spent within the degraded region.

Formally, the function η_J may incorporate both the individual values of service intensities $\tilde{\mu}_i^J$ and the geometric characteristics of the distance from the target state q^o. In this case, the model assumes that more distant states entail higher losses, all other conditions being equal. Within the framework of hybrid analysis, these values can be computed either individually for each edge node E_i or in aggregate form, using weight coefficients that reflect QoS-related priorities. The resulting value η_J is then used to refine the integral efficiency criterion by replacing the general value φ in formula $\eta = \varphi\pi(Y)$ with the corresponding localised value that models a specific risk:

$$\eta = \left(\sum_{J=1}^{c_Z} \omega_J \eta_J \right) \pi(Y) \tag{6}$$

,

where $\omega_J \in [0,1]$ is a weighting coefficient representing the significance of scenario J in the overall assessment of subsystem controllability. These coefficients may be determined based on the criticality of resources in the corresponding state or empirically, according to the likelihood of the respective configuration occurring due to typical disturbances (e.g., overload within an IoT subsegment with limited communication bandwidth).

To deepen the analysis of the dynamics of the edge-IoT subsystem under variable load conditions, we consider an alternative to the model previously presented in the section. This alternative model is based on a Markov process generator constructed using refined service probabilities and updated routing mechanisms. Such a refined model is capable of capturing threshold-based or reactive behaviour of the subsystem, in which edge nodes switch to an altered mode of operation upon reaching local overload conditions.

Let $m,n \in B$, where B is the set of indices corresponding to admissible state configurations, and $i,j \in I$, where I is the set of indices of edge nodes in the subsystem. We formalise the general generator matrix $\hat{A} = (\hat{a}_{mn})$, whose elements are defined by expressions

$$\hat{a}_{mn} = \varepsilon\left(s_i^{(m)}\right)\mu_i\theta_{ij}, \quad m \neq n, \quad i \neq j, \tag{7}$$

$$\hat{a}_{nn} = -\sum_{i=1}^{L}\varepsilon\left(s_i^{(n)}\right)\mu_i, \tag{8}$$

where $s_i^{(m)}$ denotes the number of requests at node E_i in state m, μ_i represents the service intensity at node E_i, and $\theta_{ij} \in [0,1]$ is the routing probability from node E_i to node E_j. The function $\varepsilon\left(s_i^{(m)}\right)$ quals 1 if $s_i^{(m)} > 0$ holds, and 0 otherwise.

The transition probability matrix in this model is defined by the classical matrix exponential of the generator $\hat{P}(t) = \exp\left(\hat{A}t\right)$, which describes the temporal evolution of the subsystem under the new service structure. The constructed matrix $\hat{A}$ can be used for comparative analysis with other generators, in particular $\widetilde{A^J}$, which incorporated hybrid parameters.

To specify the behaviour of the edge-IoT subsystem under particular crisis scenarios, we introduce a specialised generator matrix $\tilde{A}^J = \left(\tilde{a}_{mn}^J\right)$, defined analogously to (7) and (8), but incorporating hybrid service intensities as given in

$$\tilde{a}_{mn}^J = \varepsilon\left(s_i^{(m)}\right)\mu_{ij}^J v_{ij}^{(m)}, \quad m \neq n, \quad i \neq j, \tag{9}$$

$$\tilde{a}_{nn}^J = -\sum_{i=1}^{L}\varepsilon\left(s_i^{(n)}\right)\mu_i^J, \tag{10}$$

where $\tilde{\mu}_i^J$ denotes the hybrid service intensity at node E_i under the scenario indexed by J, and $v_{ij}^{(m)} \in [0,1]$ represents the adapted routing in state m.

The transition matrix in this case takes the form $\tilde{P}^J(t) = \exp\left(\tilde{A}^J t\right)$ and enables the evaluation of subsystem behaviour under the corresponding crisis regime. For the purpose of analysing the probability of reaching a critical state $q^{(c_\gamma + J)}$ at time t, an indicator vector $\tilde{q}^J \in \mathbb{R}^{c_Q}$ is introduced, where $\tilde{q}_n^J = 1$ if $n = c_\gamma + J$, and 0 otherwise. Using $\tilde{q}^J$, the probability function for reaching the specified scenario is defined by expression

$$F_J(t) = \tilde{q}^J \tilde{P}^J(t) e_1 \quad , \tag{11}$$

where e_1 is the unit vector corresponding to the initial state.

The combined construction enables, on the one hand, the modelling of reactive dynamics (via $\hat{A}$), and on the other – the specification of the crisis scenario through $\tilde{A}^J$, integrating both aspects into the formulation of cost functions and efficiency criteria, in particular (5) and (6).

The refined model enables the specification of loss estimates and performance characteristics of the edge-IoT subsystem under stochastic dynamics. Based on the previously formulated probability function of reaching a critical state, we formalise a generalised cost function that accounts for both the temporal aspect of the expected response and the risk of delay in transitioning to an admissible configuration. Let $F_J(t)$ denote the probability function of reaching a degraded state indexed by $J \in \{1,\ldots,c_Z\}$ before time t, and $\varphi \in \mathbb{R}_+$ be the fixed guaranteed response time threshold. Then the expected losses within scenario J are defined by expression

$$\eta_J^* = \varphi \int_0^{\varphi} t dF_J(t) + \varphi\left[1 - F_J(\varphi)\right] \quad , \tag{12}$$

which reflects the average temporal contribution to losses under the condition that the subsystem responds within the allowed interval, as well as a penalty for exceeding the admissible threshold.

2.3. Evaluation of Control Efficiency and QoS Indicators in the Edge-IoT Subsystem

Considering the Markovian structure of admissible and critical states, the expected number of requests at an edge node E_i in the stationary mode is described by an integral convolution over all admissible configurations, as expressed in

$$s_i^* = \sum_{n \in B} s_i^{(n)} \pi_n \quad , \tag{13}$$

where $s_i^{(n)}$ denotes the number of requests at node i in configuration $q^{(n)}$, and π_n represents the probability of the subsystem residing in this configuration under stationary conditions. Accordingly, each estimate (13) accounts for both the admissible domain Y and the set of states with reduced performance Z.

The probability that exactly k requests are present at the edge node i is defined by expression

$$\mathrm{P}\{s_i = k\} = \sum_{n \in B, s_i^{(n)} = k} \pi_n \quad , \tag{14}$$

which ensures the correct aggregation of the probability distribution over the hyperplane of configurations with a fixed level of local load. In this case, the queue is not considered in isolation but within the context of the global state of the subsystem as a whole.

The estimate of the average effective arrival rate of requests to the edge node E_i, considering all states of the subsystem, is formulated as in

$$\lambda_i = \sum_{n \in B} \varepsilon\left(s_i^{(n)}\right) \mu_i \pi_n \quad (15)$$

,

where $\varepsilon\left(s_i^{(n)}\right)$ is an indicator of the presence of requests at node i in state $q^{(n)}$, and μ_i denotes the service rate of requests under normal operating conditions. If the edge-IoT subsystem is in a critical state $q^{(n)} \in Z$, it is advisable to use the hybrid rate $\tilde{\mu}_i^J$ instead of μ_i, defined over the corresponding segment. This approach enables the incorporation of reactive changes in the service depending on the current state.

The average time a request spends at the edge node, which reflects the relationship between the load and the service rate, is given by the expression

$$u_i^* = s_i^* / \lambda_i \quad (16)$$

.

This represents a generalisation of Little's law for a Markovian system with a set of controlled and degraded configurations. Expression (16) retains its interpretation as a characteristic of request delay at the level of an individual node but incorporates a global account of system states.

Finally, the load coefficient of the edge node is defined by the expression

$$\psi_i = \lambda_i / \mu_i \quad (17)$$

,.

which reflects the expected proportion of time during which the edge node E_i operates in service mode. This value serves as a key QoS indicator within the edge-IoT subsystem employing hybrid routing, as it enables the identification of nodes prone to overload or delays along routing paths. In cases where $\tilde{\mu}_i^J$ is known, the coefficient ψ_i may be refined by incorporating hybrid load parameters.

3. Results and Discussion

To verify the effectiveness of the proposed hybrid Markov model of controlled load distribution in the edge-IoT subsystem, a computer simulation was carried out using the MATLAB R2023a environment. A subsystem consisting of four edge nodes was modelled, each represented by a classical M/M/1 queueing system enhanced with mechanisms for adaptive regulation of service intensity under varying load conditions. The incoming traffic flows were characterised by parameters approximating an exponential distribution, with an average intensity $\lambda_k \in [0.7, 1.3]$, which allowed for the consideration of different operational modes, including overload, partial resource degradation, and changes in topological availability.

The generation of the set of admissible states of the subsystem was carried out by iterating over all queue vectors $\vec{q}(n) \in \mathbb{Z}_+^4$ that satisfied the condition of Euclidean proximity to the target vector $q^o = (2,2,2,2)$, with an accuracy defined by the deviation $\left\| q(n) - q^o \right\|_\infty \leq 1$. This approach enabled the construction of the set Q, consisting of 81 elements, which was further divided into the set of admissible states Y and the critical subset $Z = Q \setminus Y$, in accordance with the individual tolerance vectors $b = (1,1,1,1)$. Each edge node was assigned an individual service rate μ_k, which varied within the bounds of $[0.8, 1.2]$ and reflected the heterogeneous performance of devices in scenarios characterised by energy or computational constraints.

To model the probabilistic dynamics of transitions, a stochastic routing matrix $\Theta = \left[\theta_{ij} \right]$ was employed, in which each element represented the proportion of requests redirected from node j to node i. These coefficients were selected to ensure, on average, a balanced load within the permissible deviation. Based on the obtained parameters, a Markov process generator A^J was constructed, where for each pair of states (m,n), the transition matrix elements were numerically computed using formulas that incorporated a hybrid estimate of the local intensity $\tilde{\mu}_k$. In critical states from the set Z, this estimate was modified by accounting for the dynamic parameters h_i and υ_i, which reflected both the actual queue length and the structural routing, formally expressed as $\upsilon_i = \sum_{j=1}^{L} \theta_{ij} \mu_j$.

The computation of the matrix exponential $\exp\left(A^J t \right)$, which describes the temporal evolution of the probabilities of the system being in a given state, was performed using the expm function from the MATLAB Symbolic Math Toolbox, as numerical stability was of critical importance for the validation of policies under varying parameters $t \in [0,50]$. The indicator function for reaching each of the critical states was implemented by constructing a vector $q^{(J)}$, in which only a single component assumed the value 1. The calculation of the hybrid cost function was carried out in accordance with expression (5) via numerical integration using the integral function, which ensured high-precision approximation of the result at small values of φ.

The stationary distribution $\pi = (\pi_n)$ was obtained as the solution to the system of Kolmogorov equations, supplemented by the normalisation condition $\sum_n \pi_n = 1$. The solution was computed using the linsolve function from the MATLAB Linear Algebra Toolbox, with the Rectangular option enabled to handle a system with a degenerate extended dimension. The integral probability of the system remaining in admissible states $\pi(Y)$ was calculated by summing the components of the stationary distribution corresponding to the indices belonging to the set Y. Finally, the generalised performance criterion was computed as the product $\eta = \varphi \pi(Y)$, providing a quantitative interpretation of the stability and quality of the routing policy, taking into account the potential loss of QoS due to perturbations or non-uniformity of incoming traffic.

Based on the constructed model, an experimental study of the dynamics of the edge-IoT subsystem was conducted over the time interval $t \in [0,50]$, during which the subsystem evolved according to a Markov process with an adaptive routing policy. The key indicator of dynamic behaviour was the generalised performance criterion $\eta(t) = \varphi \pi(Y,t)$, which reflects the integral probability of the subsystem remaining within admissible configurations, taking into account a guaranteed stabilisation threshold. The value $\pi(Y,t)$ was computed as the sum of the components of vector $\pi(t) = e_0 \exp\left(A^J t \right)$, where e_0 denotes the unit vector of the initial state.

To compare the effectiveness of the adaptive model, two baseline strategies were also simulated – round-robin redirection and random request distribution. In the former case, each request was sequentially redirected to the next node, with a fixed transition probability, ensuring uniform load distribution but failing to account for the current queue states. In the latter case, a fully stochastic principle was implemented, whereby routes were selected randomly without consideration of the load history.

Fig. 1 presents the results of computing the criterion $\eta(t)$ for the three considered approaches. The adaptive strategy exhibits a stable increase in performance from the early stages of the simulation, reaching the value $\eta(t) \approx 0.93$ already at $t \approx 20$, and approaching the asymptotic level $\eta_\infty \approx 0.95$. This indicates the rapid entry of the edge-IoT subsystem into the stability zone and a high reliability in maintaining QoS. In contrast, the round-robin policy proved inertial – its $\eta(t)$ value increased more slowly, approaching only 0.85. The random redirection strategy was the least effective: due to excessive transition dispersion, the subsystem spent a considerable portion of time in configurations from the set Z, leading to a significant degradation in QoS.

Transitioning from the dynamic assessment of subsystem stabilisation to a quantitative comparison of control strategies, the focus now shifts to the experimental results, which illustrate how different routing policies affect key performance indicators, in particular the average queueing delay, resource utilisation level, and the value of the cost function.

The queueing delay was calculated by recording the arrival time of each request and registering the completion time of its service at the corresponding edge node. The average value was computed over the time interval $t \in [0,50]$, within which the simulation was conducted. The utilisation coefficient was defined as the ratio of the total server busy time to the overall observation time, with values aggregated across all nodes of the edge-IoT subsystem. The cost function $\Psi(n)$ was computed for each subsystem configuration as a weighted sum of queue lengths, taking into account the weighting coefficients α_k, and subsequently averaged over all admissible states.

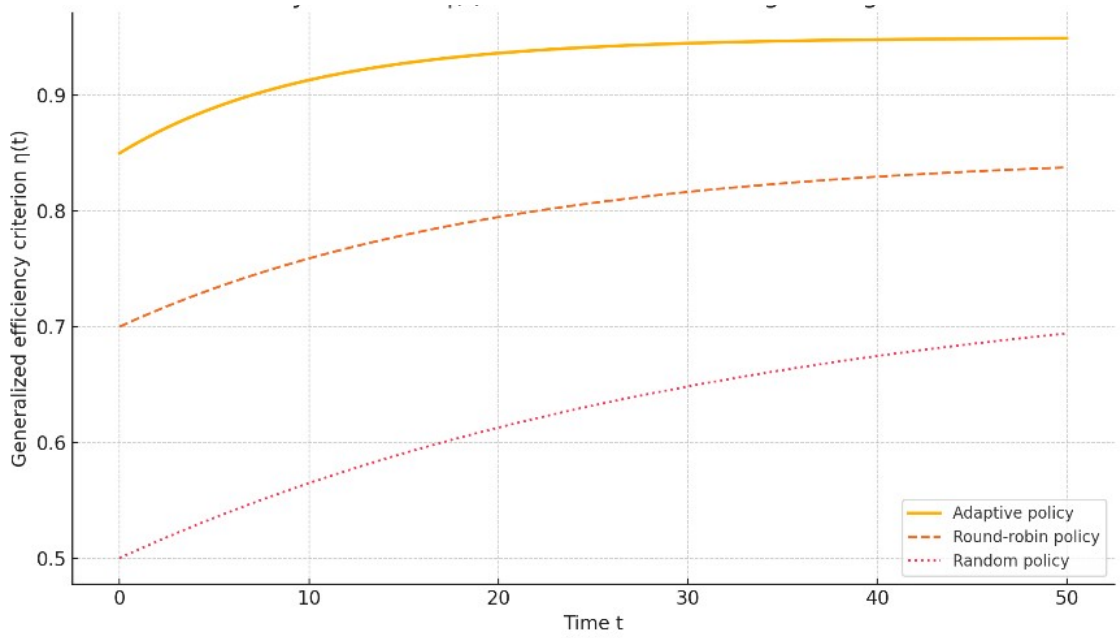

Figure 1: Dynamics of η(t) under different routing strategies.

The results are presented as three comparative diagrams shown in Fig. 2. The first diagram demonstrates that the average queueing delay under the adaptive strategy is approximately 1.2 seconds, whereas for the round-robin and random approaches, these values reach 2.3 and 3.7 seconds, respectively. This difference indicates a more efficient load balancing within the set of permissible states, achieved through dynamic routing. The second diagram illustrates that adaptive control leads to an average edge node utilisation of 92%, while the baseline policies show lower performance – 85% and 78%, respectively. This metric is particularly significant in the context of

energy consumption, as it reflects the effective use of available computational resources without excessive idle time.

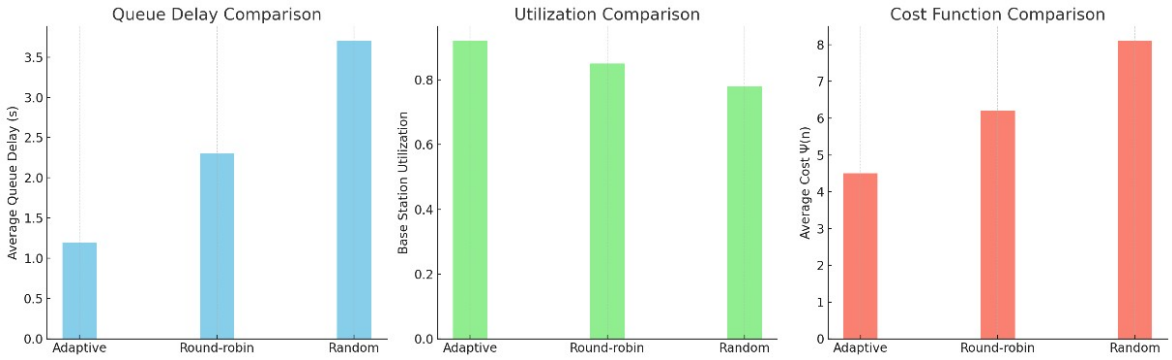

Figure 2: Quantitative evaluation of delays, utilisation, and costs for the three routing strategies.

Finally, the third diagram compares the values of the cost function $\Psi(n)$. The adaptive strategy reduces the average value to 4.5 arbitrary units, whereas the round-robin and random strategies yield values of 6.2 and 8.1, respectively. This confirms the capability of the adaptively controlled subsystem to maintain the load configuration closer to the optimal one, which was formalised in the model section by the vector q^o and the set Y. It is worth noting that all the values presented were computed in the MATLAB environment using scripts based on the functions mean, sum, find, and accumarray, which enabled efficient aggregation of simulation results across many configurations.

The analysis of the experimental results enables the formulation of several conceptual and applied conclusions regarding the efficiency of various routing approaches in edge-IoT subsystems. The adaptive policy, developed based on a Markovian model that accounts for the current state of the subsystem and admissibility criteria, demonstrated a clear advantage over classical schemes – both in terms of integral metrics and the phase dynamics of state transitions.

Firstly, consistently lower delay values alongside a high node utilisation ratio indicate the ability of the adaptive model to balance the load effectively, avoiding both overload and inefficient resource idleness. This directly influences the level of QoS, which remains stable even under varying traffic conditions. Secondly, the reduction in the average value of the cost function $\Psi(n)$ suggests that the adaptively managed subsystem operates closer to the target load distribution, thereby minimising internal losses related to routing, energy consumption, and processing time.

The phase analysis further confirmed that the adaptive strategy not only enables the attainment of a stable operating regime but also does so more rapidly and with smaller fluctuations compared to round-robin or random routing. Given the set of admissible configurations Y introduced in the model, the adaptive policy ensures not only the reachability of the desired state but also sustained operation within it for a significant period, thereby providing long-term stability of the edge-IoT subsystem.

From an applied perspective, the proposed model can be implemented as a foundation for the development of control mechanisms in real-world edge-IoT subsystems, particularly within the context of distributed intelligence in 6G networks. Its advantages are especially evident in scenarios characterised by unpredictable traffic fluctuations, dynamic topology, or partial resource unavailability. Moreover, the formalised efficiency criteria (in particular, the integral metric $\eta = \varphi\pi(Y)$) can be adapted to systems with alternative types of constraints or priorities, thereby broadening the applicability of the proposed approach.

Conclusions

In modern edge-IoT subsystems, which combine high input traffic dynamics, structural heterogeneity of nodes, and strict requirements for maintaining stable service quality, there arises a need for the development of formalised models capable of ensuring real-time load controllability. The motivation for this study stems from the necessity to design an adaptive model that, on the one hand, reflects the stochastic nature of subsystem behaviour and, on the other, provides mathematically grounded system stabilisation within admissible bounds, even under conditions of disturbances and partial resource degradation.

The scientific novelty of the present work lies in the development of a hybrid Markovian model for controlled load distribution in edge-IoT subsystems, which for the first time integrates a structured set of admissible states with parameterised transition probabilities between them, formalised through the generator of a stochastic process incorporating adaptive routing. Unlike existing analogues, this model introduces an integral efficiency criterion that combines the probability of the system residing in QoS-defined states with an estimate of the guaranteed stabilisation time, while also accounting for localised losses incurred upon entering critical configurations. Additionally, a mechanism is proposed for constructing a routed policy that not only minimises the average service cost but also ensures subsystem resilience to fluctuations in traffic, topology structure, and node performance. This fundamentally distinguishes the approach from heuristic or metaheuristic algorithms that lack a clearly defined spatial dynamic of the system.

The practical value of the developed model lies in its potential implementation as a foundation for designing control mechanisms within distributed edge-IoT architectures, particularly in scenarios where service continuity is critical – for instance, in autonomous transport systems, eHealth applications, defence-oriented sensor networks, and energy-autonomous microsystems. The integral controllability metric, formulated as the product of the stabilisation probability and the stationary probability of remaining within the admissible states, can serve as a universal criterion for load optimisation in the design of adaptive controllers operating in environments with partially observable parameters.

The synthesis of the computer simulation results confirmed the effectiveness of the proposed strategy in comparison with classical approaches: the average delay was reduced by nearly a factor of three, resource utilisation was increased to 92%, and the cost function was lowered to 4.5 a.u. The adaptive strategy demonstrates rapid convergence to the stability zone and maintains QoS even under conditions of intensive load disturbances, which substantiates its practical reliability in real-world environments.

Future research may be directed towards extending the model to support multicluster IoT architectures with dynamic topology, incorporating non-Poisson traffic sources, applying reinforcement learning models to develop reactive policies, and analysing subsystem behaviour under attack scenarios or deliberate performance degradation. These directions open the prospect for developing robust and interpretable solutions applicable across a broad range of mission-critical applications in the field of intelligent distributed computing.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish its results.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Al-Sarawi, S., Anbar, M., Abdullah, R., and Al Hawari, A. B., Internet of Things Market Analysis Forecasts, 2020–2030, 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) 2020 449–453. doi:10.1109/worlds450073.2020.9210375.

[2] Zaman, M., Puryear, N., Abdelwahed, S., and Zohrabi, N., A Review of IoT-Based Smart City Development and Management, Smart Cities 2024 1462–1501. doi:10.3390/smartcities7030061.

[3] Cao, K., Liu, Y., Meng, G., and Sun, Q., An Overview on Edge Computing Research, IEEE Access 2020 85714–85728. doi:10.1109/access.2020.2991734.

[4] Laroui, M., Nour, B., Moungla, H., Cherif, M. A., Afifi, H., and Guizani, M., Edge and fog computing for IoT: A survey on current research activities & future directions, Computer Communications 2021 210–231. doi:10.1016/j.comcom.2021.09.003.

[5] Lone, K., and Sofi, S. A., A review on offloading in fog-based Internet of Things: Architecture, machine learning approaches, and open issues, High-Confidence Computing 2023 100124. doi:10.1016/j.hcc.2023.100124.

[6] Mehmood, M. Y., Oad, A., Abrar, M., Munir, H. M., Hasan, S. F., Muqeet, H. A. ul, and Golilarz, N. A., Edge Computing for IoT-Enabled Smart Grid, F. Lombardi (Ed.), Security and Communication Networks 2021 1–16. doi:10.1155/2021/5524025.

[7] Apat, H. K., Sahoo, B., Goswami, V., and Barik, R. K., A hybrid meta-heuristic algorithm for multi-objective IoT service placement in fog computing environments, Decision Analytics Journal 2024 100379. doi:10.1016/j.dajour.2023.100379.

[8] Latip, R., Aminu, J., Hanafi, Z. M., Kamarudin, S., and Gabi, D., Metaheuristic task offloading approaches for minimization of energy consumption on edge computing: a systematic review, Discover Internet of Things 2024. doi:10.1007/s43926-024-00089-y.

[9] Kiani, F., and Seyyedabbasi, A., Metaheuristic Algorithms in IoT: Optimized Edge Node Localization, Studies in Computational Intelligence 2022 19–39. doi:10.1007/978-3-031-16832-1_2.

[10] Lebovka, N., Petryk, M., Tatochenko, M. and Vygornitskii, N., Two-stage random sequential adsorption of discorectangles and disks on a two-dimensional surface, Physical Review E 2023 108, 024109. doi: 10.1103/PhysRevE.108.024109

[11] Lebovka, N., Petryk, M., Vorobiev, E, Monte Carlo simulation of dead-end diafiltration of bidispersed particle suspensions, Physical Review E 2022 106 064610. doi: 10.1103/PhysRevE.106.064610

[12] Bhardwaj, K. K., Banyal, S., Sharma, D. K., and Al-Numay, W., Internet of things based smart city design using fog computing and fuzzy logic, Sustainable Cities and Society 2022 103712. doi:10.1016/j.scs.2022.103712.

[13] Shi, Y., Chu, J., Ji, C., Li, J., and Ning, S., A Fuzzy-Based Mobile Edge Architecture for Latency-Sensitive and Heavy-Task Applications, Symmetry 2022 1667. doi:10.3390/sym14081667.

[14] Abdulazeez, D. H., and Askar, S. K., A Novel Offloading Mechanism Leveraging Fuzzy Logic and Deep Reinforcement Learning to Improve IoT Application Performance in a Three-Layer Architecture Within the Fog-Cloud Environment, IEEE Access 2024 39936–39952. doi:10.1109/access.2024.3376670.

[15] Qafzezi, E., Bylykbashi, K., Ampririt, P., Ikeda, M., Matsuo, K., and Barolli, L., An Intelligent Approach for Cloud-Fog-Edge Computing SDN-VANETs Based on Fuzzy Logic: Effect of Different Parameters on Coordination and Management of Resources, Sensors 2022 878. doi:10.3390/s22030878.

[16] Soleimanikia, M., Bushehrian, O., and Mahmoodi, D., A Novel Graph-Based Energy Efficient Sensor Selection Scheme in Edge Computing, 2023 International Conference on Smart Applications, Communications and Networking (SmartNets) 2023 1–6. doi:10.1109/smartnets58706.2023.10216179.

[17] Jiang, Z., Li, J., Hu, Q., Meng, W., Pedrycz, W., and Su, Z., Scalable Graph-Aware Edge Representation Learning for Wireless IoT Intrusion Detection, IEEE Internet of Things Journal 2024 26955–26969. doi:10.1109/jiot.2024.3397364.

[18] Lu, Z., Chang, Z., He, M., and Song, L., Zero-Shot Traffic Identification with Attribute and Graph-Based Representations for Edge Computing, Sensors 2025 545. doi:10.3390/s25020545.

[19] Christalin Nelson, S., Singh, R. K., and Prakash, G. L., Hybrid deep learning model based on Intelligent Microbat Routing (IMR) and Popularity Content Caching (PCC) for an effective caching and routing in vehicular edge networks, Computers and Electrical Engineering 2022 108353. doi:10.1016/j.compeleceng.2022.108353.

[20] Alwakeel, A. M., Enhancing IoT performance in wireless and mobile networks through named data networking (NDN) and edge computing integration, Computer Networks 2025 111267. doi:10.1016/j.comnet.2025.111267.

[21] Jiang, W., Han, H., Zhang, Y., Wang, J., He, M., Gu, W., Mu, J., and Cheng, X., Graph Neural Networks for Routing Optimization: Challenges and Opportunities, Sustainability 2024 9239. doi:10.3390/su16219239.

[22] Heidari, A., Jamali, M. A. J., Navimipour, N. J., and Akbarpour, S., A QoS-Aware Technique for Computation Offloading in IoT-Edge Platforms Using a Convolutional Neural Network and Markov Decision Process, IT Professional 2023 24–39. doi:10.1109/mitp.2022.3217886.

[23] Kalnoor, G., and S, G., Markov Decision Process based Model for Performance Analysis an Intrusion Detection System in IoT Networks, Journal of Telecommunictions and Information Technology 2021 42–49. doi:10.26636/jtit.2021.151221.

[24] Sahu, D., Nidhi, Chaturvedi, R., Prakash, S., Yang, T., Rathore, R. S., Wang, L., Tahir, S., and Bakhsh, S. T., Revolutionizing load harmony in edge computing networks with probabilistic cellular automata and Markov decision processes, Scientific Reports 2025. doi:10.1038/s41598-025-88197-9.

[25] Chen, W., Chen, Y., and Liu, J., Service migration for mobile edge computing based on partially observable Markov decision processes, Computers and Electrical Engineering 2023 108552. doi:10.1016/j.compeleceng.2022.108552.

[26] Skirelis, J., and Navakauskas, D., Edge computing in IoT: Preliminary results on modeling and performance analysis, 2017 5th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE) 2017 1–4. doi:10.1109/aieee.2017.8270555.