UDC 004.9

PREDICTIVE MULTI-LAYER RESOURCE SLICING FOR 5G NETWORKS

Sulima S. V. – PhD, Associate Professor, Associate Professor of the Department of Information Technologies in Telecommunications, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0000-0002-6333-7693.

Karashevych Ie. D. – Post-graduate student at the Department of Information Technologies in Telecommunications, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0009-0000-0671-8185.

ABSTRACT

Context. The rapid deployment of 5G networks and the emergence of 6G architectures introduce unprecedented traffic heterogeneity and burstiness across radio, edge, and core domains. Meanwhile, the energy footprint of mobile infrastructure is becoming a major sustainability concern, as carbon emissions increasingly shape network operation policies.

Objective. This work aims to design a predictive carbon-aware multi-layer resource slicing framework for RAN – edge – core 5G/6G networks that jointly optimizes latency, cost, energy, and carbon emissions under bursty traffic conditions.

Method. The proposed approach integrates an M/G/1-based queuing model for accurate representation of heavy-tailed service times and bursty arrival patterns; hybrid short-term/long-term forecasting of both traffic load and regional carbon intensity; and multi-objective optimization for carbon-aware VNF placement and traffic steering across network layers. A proactive – reactive orchestration mechanism performs predictive resource pre-allocation and runtime scaling.

Results. Trace-driven simulations on a representative multi-layer testbed demonstrate a 34% reduction in CO_2 emissions compared to latency-first orchestration, alongside a 22% decrease in operational cost and <1% SLA violation rate. Tail latency remains within slice-specific thresholds even under bursty loads, confirming that carbon reductions can be achieved without service degradation.

Conclusions. Predictive, carbon-aware orchestration across RAN-edge-core domains substantially improves environmental and economic efficiency while preserving QoS guarantees. The results highlight the importance of integrating forecast-driven optimization and realistic traffic modeling into next-generation slicing architectures.

KEYWORDS: carbon-aware networking, multi-layer slicing, 5G/6G, predictive orchestration, bursty traffic, M/G/1 queueing, edge-cloud, sustainable networking.

ABBREVIATIONS

API is an Application Programming Interface;

ARIMA / SARIMA is a (Seasonal) AutoRegressive Integrated Moving Average;

CI is a Carbon Intensity;

CPU is a Central Processing Unit;

DR is a Demand Response;

DRU is a Distributed Radio Unit;

eMBB is an Enhanced Mobile Broadband;

EWMA is an Exponentially-Weighted Moving Average;

GBM is a Gradient Boosting Machine;

LSTM is a Long Short-Term Memory (neural network);

MEC is a Multi-access Edge Computing;

mMTC is a Massive Machine-Type Communications;

NFV is a Network Function Virtualization;

NSI is a Network Slice Instance;

NSSI is a Network Slice Subnet Instance;

OPEX is an Operational Expenditure;

QoS is a Quality of Service;

RAN is a Radio Access Network;

SDN is a Software-Defined Networking;

SFC is a Service Function Chain;

SLA is a Service-Level Agreement;

URLLC is an Ultra-Reliable Low-Latency Communications;

VNF is a Virtual Network Function.

NOMENCLATURE

α is a share of electricity from source i;

α is an exponent in the nonlinear power model ($P \propto r^{\alpha}$);

$\alpha(t)$ is a share of electricity from source i at time t;

B is a bandwidth capacity of node j;

$b_{j,k}(t)$ is a bandwidth allocated to slice k on node j at time t;

$Carbon(t)$ is a total CO_2 emissions at time t;

$CI_n(t)$ is a carbon intensity of node n at time t;

C_j is a computational capacity of node j;

c_j^{comp} is a computation unit cost on node j;

c_j^{net} is a bandwidth unit cost on node j;

C_s^2 is a coefficient of variation of service time;

$Cost(t)$ is a total operational cost at time t;

D_k^{prp} is a propagation delay for slice k;

D_k^{proc} is a processing delay;

$D_k^{current}$ is a current end-to-end delay of slice k;

$E[S]$ is an expected (mean) service time;

$E[S^2]$ is a second moment of service time;

$Energy(t)$ is a total energy consumption at time t;

DOI 10.15588/1607-3274-2026-2-2

$f_k(e \to c)(t)$ is a fraction of slice k traffic offloaded from edge to core at time t;

J is an objective function value (cost + energy + carbon + latency);

$Latency(t)$ is an aggregated latency at time t;

$L_k^{\max}$ is a maximum allowable latency (SLA constraint) for slice k;

$MAPE$ is a Mean Absolute Percentage Error;

μ_k is a mean service rate for slice k;

μ is a memory requirement of VNF_i in slice k;

P is a set of feasible deployment layers for VNF_i of slice k;

P_j^{idle} is an idle power consumption of node j;

$P_j(t)$ is a power usage of node j at time t;

π_k is a priority score of slice k;

$Q_j(t)$ is a queue length of node j at time t;

$Q^{\max}$ is a maximum allowable queue length;

$r_j(t)$ is a total compute resources allocated on node j;

$r_{j,k}(t)$ is a compute resources of node j allocated to slice k;

S_i^k is a service time of VNF_i in slice k;

S_j is an available memory capacity of node j;

SFC_k is a service function chain of slice k;

τ is a prediction error threshold;

$Var[S]$ is a variance of service time;

VNF is a virtual network function i of slice k;

w_1 is a weight coefficient for cost objective;

w_2 is a weight coefficient for energy objective;

w_3 is a weight coefficient for carbon objective;

w_4 is a weight coefficient for latency objective;

W_k is an average queue waiting time for slice k;

$x_{i,j}^k(t)$ is a binary decision variable indicating placement of VNF_i of slice k on node j.

INTRODUCTION

The proliferation of 5G networks and the forthcoming 6G era [1] promise unprecedented data rates, ultra-low latency, and massive connectivity. Resource slicing, also known as network slicing, is a pivotal innovation in 5G core networks that allows for the creation of multiple virtual networks or "slices" operating on a single physical infrastructure. This technology is essential for addressing diverse application requirements – such as security, reliability, and performance – enabling various industries to harness tailored connectivity solutions that can enhance operational efficiency and user experience. As enterprises increasingly rely on advanced mobile networking, particularly with the deployment of 5G Standalone (5G SA), resource slicing has emerged as a critical feature, accommodating the specific needs of applications that demand guaranteed latency, speed, and connection density [3]. Utilizing technologies like Software-Defined Networking (SDN) and Network Function Virtualization (NFV), resource slicing allows network operators to dynamically allocate resources and manage multiple service levels concurrently. The benefits of network slicing are manifold, including improved resource utilization, customizable services, cost-effectiveness, and enhanced security measures. Despite the clear advantages, the deployment of network slicing necessitates careful orchestration and management to maintain service quality while addressing evolving security threats and integration challenges [4].

This technological advancement comes at a significant environmental cost. Mobile network operators consume approximately 2–3% of global electricity, with projections indicating exponential growth as traffic demands surge. As network infrastructure expands to support diverse use cases – the carbon footprint of telecommunications becomes a critical concern.

Traditional resource allocation strategies primarily optimize for performance metrics such as latency, throughput, and resource utilization. While recent works have explored energy efficiency, few systematically integrate carbon awareness across the entire RAN-edge-core continuum. Moreover, the inherent burstiness of modern network traffic – characterized by heavy-tailed distributions and self-similar properties – poses fundamental challenges for resource provisioning and carbon optimization.

The object of study is the multi-layer 5G/6G communication infrastructure, which includes the RAN, edge, and core domains participating in end-to-end network slicing and service provisioning.

The subject of study is the predictive and carbon-aware resource slicing/orchestration mechanisms (models, algorithms, and optimization strategies) that dynamically allocate compute and network resources under bursty traffic.

The purpose of the work is to develop a predictive, carbon-aware multi-layer slicing framework that minimizes carbon emissions and operational cost while preserving latency and SLA guarantees under real-world bursty traffic conditions.

1 PROBLEM STATEMENT

Existing research, such as the ACNRA algorithm [5], has shown the benefits of topology- and energy-aware slicing by minimizing resource cost and energy consumption in the 5G core. However, three major challenges remain:

1. Limited traffic modeling: current approaches rely on simplified M/M/1 queueing, which does not capture burstiness and heavy-tailed inter-arrival times observed in real 5G traffic.

DOI 10.15588/1607-3274-2026-2-2

2. Neglect of carbon footprint: energy-aware models minimize consumption but ignore the carbon intensity of energy sources.

3. Lack of multi-layer orchestration: slicing decisions are made independently at core or edge levels, leading to suboptimal end-to-end performance.

To address these challenges, we propose a Predictive Carbon-Aware Multi-Layer Resource Slicing algorithm:

1. Hybrid queuing framework: We develop a novel M/G/1-based analytical model that accurately captures bursty traffic characteristics through general service time distributions while maintaining tractability for online optimization.

2. Predictive carbon-aware orchestration: A machine learning pipeline predicts both traffic patterns and regional carbon intensity, enabling proactive workload placement and resource scaling decisions across RAN, edge, and core layers.

3. Multi-objective optimization: We formulate a constrained optimization problem that simultaneously minimizes operational cost, energy consumption, carbon emissions, and end-to-end latency while respecting slice-specific SLA constraints.

We consider a three-tier 5G/6G network architecture.

RAN Layer (L_1): Distributed radio units (DRUs) and baseband processing units handle radio resource allocation and initial traffic aggregation. Each DRU $i \in \Re$ serves a geographical cell with time-varying traffic load $\lambda(t)$.

Edge Layer (L_2): MEC servers $\{E_1, E_2, ..., E\}$ deployed at aggregation points provide low-latency processing for latency-sensitive applications. Each edge node E_j has computational capacity C_j^{edge} (in CPU cores), storage S_j^{edge}, and network bandwidth B_j^{edge}.

Core Layer (L_3): Centralized cloud datacenters $\{D_1, D_2, ..., D_n\}$ offer high computational capacity for processing-intensive tasks with relaxed latency requirements.

Network slices are realized through service function chains (SFCs), where each slice type $k \in K$ requires a sequence of virtual network functions (VNFs):

$$SFC_k = < VNF_1^k \to VNF_2^k \to ... \to VNF_{n_k}^k >.$$

Each VNF_i^k is characterized by:

– Computational demand: ρ_i^k (CPU cycles per request);

– Memory footprint: μ_i^k (GB);

– Processing time distribution: Service time follows a general distribution $G(\cdot)$ with mean $E[S_i^k]$ and variance $Var[S_i^k]$;

– Placement constraints: $\rho_i^k \subseteq \{L_1, L_2, L_3\}$ indicates feasible deployment layers.

To capture traffic burstiness, we model each VNF instantiation as an M/G/1 queue:

Arrival Process: Requests arrive following a Poisson process with time-varying rate $\lambda_k(t)$, which is decomposed into predictable component $\chi_k(t)$ obtained via LSTM-based forecasting and a stochastic component $\varepsilon_k(t) \sim N(0, \sigma_k^2)$.

Service Process: Service times follow a general distribution with:

– Mean: $E[S] = 1/\mu_k$;

– Second moment: $E[S^2]$;

– Coefficient of variation: $C_s^2 = Var[S]/([S])^2$.

By the Pollaczek-Khinchine formula, the mean queue waiting time is:

$$W_k = \frac{\lambda_k E[S^2]}{2(1-\rho_k)},$$

where $\rho_k = \lambda_k / \mu_k$ is the utilization factor and $E[S^2] = (E[S])^2(1 + C_s^2)$.

For heavy-tailed service distributions (e.g., Pareto), high C_s^2 values significantly impact latency, necessitating careful resource provisioning.

Carbon intensity (CI) quantifies grams of CO_2 emitted per kilowatt-hour of electricity consumed. For each node n at time t :

$$CI_n(t) = \sum \alpha_i(t) \cdot e_i ,$$

where $\alpha(t)$ is the fraction of electricity from source i (coal, natural gas, solar, wind, etc.) and e is its emission factor (gCO_2/kWh).

We employ a hybrid prediction model. Short-term (1–6 hours): Gradient boosting on weather forecasts, historical generation data. Long-term (6–24 hours): ARIMA with seasonal decomposition

2 REVIEW OF THE LITERATURE

The architectural foundation of 5G core network slicing builds upon virtualization technologies that enable the partitioning of physical infrastructure into multiple logical networks. Infrastructure sharing through slicing relies on virtualization and isolation of compute, storage, and network resources, where Virtual Machines (VMs) allow slicing of compute and storage resources (e.g. CPUs and RAM), while Virtual Networks (VNs) allow the slicing on network links and nodes (e.g. wavelength, frequency, bandwidth, transponders) [6].

The standardization bodies have established a hierarchical model for network slice architecture. Based

DOI 10.15588/1607-3274-2026-2-2

on this layered model, 3GPP released the standard 3GPP TR 28.801, which defines a Network Slice Instance (NSI) as a set of Network Slice Subnet Instances (NSSI), composed by at least one Virtual Network Function (VNF) and/or Physical Network Functions (PNFs).

At the operational level, each network slice can have its own network architecture, including on-demand service application, resource capacity, and control policy, to balance the disparate requirements between heterogeneous services [7]. The 5G core (5GC) is designed to be "cloud native" where NFV is leveraged to create network slices, with a 5G core slice composed of a collection of 5G core VNFs that are chained together to support a specific use case [8].

Network slicing operates through a multi-tier resource allocation approach, where different tenants share the same communications and computing resources, with each virtual network allocated a certain amount of resources and then re-allocating them to its subscribers based on specific rules [9].

The evolution of network slicing architecture has progressed significantly from its inception as Dedicated Core Network (DCN) in 4G-LTE to its current state in 5G-Advanced, with enhancements categorized into phases: 5G-Basic (Release 15), early 5G-Evolution (Release 16), and advanced 5G-Evolution (Release 17) [10]. This architectural evolution supports both vertical slicing (segmentation of the core mobile network domain based on predefined applications) and horizontal slicing (designed to accommodate scaling system capacity and support computation offloading to the network edge) [11].

Recent advances in 5G resource slicing can be grouped into three categories:

Optimization-based methods (e.g., ACNRA [5], ARA [12]) – provide accurate modeling but assume static or Poisson traffic and focus solely on the core.

Machine-learning-based methods (e.g., TORCH [13], GreenSFC [14]) – enable adaptability but suffer from high training cost, poor scalability, and lack of carbon awareness.

Green and sustainable networking approaches (ITU, 2024; ETSI ENI, 2023) – address CO_2 optimization in data centers, but not in end-to-end mobile networks.

Existing approaches suffer from three key limitations: absence of a multi-layer slicing model spanning RAN – Edge – Core; lack of integration between QoS, energy, and carbon objectives; simplified queueing models ignoring bursty traffic patterns: no predictive adaptation to traffic and renewable fluctuations. The proposed framework fills these gaps.

3 MATERIALS AND METHODS

We consider a multi-layer 5G/6G network infrastructure consisting of three layers: Radio Access Network (RAN), Edge, and Core. Let K be the set of network slices, and each slice $k \in K$ is represented by a service function chain:

$$SFC_k = VNF_1^k \to VNF_2^k \to \ldots \to VNF_{n_k}^k > .$$

Given

1. Traffic model

For each slice k the incoming traffic rate is time-varying and defined as:

$$\lambda_k\ (t) = \chi_k(t) + \varepsilon_k(t),$$

where $\chi_k(t)$ is the predicted arrival rate and $\varepsilon_k(t)$ is a stochastic prediction error.

2. Service model

Each VNF is modeled as an M/G/1 queue with a general service time distribution characterized by mean $E[S_k]$ and second moment $E[S_k^2]$.

3. Infrastructure constraints

Each node j has limited computational, memory, and bandwidth resources (C_j, S_j, B_j).

4. Carbon intensity

Each node j at time t is associated with a carbon intensity value $CI_j(t)$, representing grams of CO^2 per kWh.

5. Quality-of-Service constraints

Each slice k has a maximum allowable end-to-end latency $L_k^{\max}$ specified by its SLA.

Decision Variables

– $x_{ijk}(t) \in 0,1$: placement variable indicating whether VNF i of slice k is deployed on node j;

– $r_{jk}(t) \geq 0$: computational resources allocated to slice k on node j;

– $f_k^{e \to c}(t) \in [0,1]$: fraction of slice traffic offloaded from edge to core.

Objective

Minimize the weighted multi-objective cost function:

$$\min J = w_1 \cdot Cost + w_2 \cdot Energy + w_3 \cdot Carbon + w_4 \cdot Latency,$$

where each term corresponds to operational cost, energy consumption, carbon emissions, and end-to-end latency, respectively.

Subject to constraints

1. Resource capacity constraints

$$\sum_{k=1}^{n} \sum_{i=1}^{m} x_{ij}^k \rho_i^k \leq C_j, \forall j\ .$$

2. VNF placement constraints

$$\sum_{j=1}^{n} x_{ij}^k = 1, \forall i,k.$$

DOI 10.15588/1607-3274-2026-2-2

3. Queue stability

$$\rho_j(t) < 1 .$$

4. SLA constraints

$$Latency_k(t) \le L_k^{\max}, \forall k.$$

The problem consists in determining the optimal VNF placement, resource allocation, and traffic offloading strategy that minimizes the objective function while satisfying all system and SLA constraints under bursty traffic conditions.

Our framework consists of four key components that work together to optimize network operations. The Traffic Predictor Module employs an LSTM network with attention mechanism trained on historical load patterns, taking as input features such as time-of-day, day-of-week, special events, and weather conditions to produce predicted arrival rates $\hat{\lambda}_k(t+\Delta)$ for planning horizons ranging from 5 minutes to 2 hours. The Carbon Intensity Forecaster uses an ensemble model that combines gradient boosting machines for short-term predictions with SARIMA for long-term forecasting, incorporating real-time grid data through API integration with services like ElectricityMap to achieve prediction accuracy with a mean absolute percentage error below 8% for 1-hour ahead forecasts. The Multi-Layer Optimizer addresses the NP-hard placement problem by decomposing it into tractable subproblems across three stages: first applying dynamic programming for VNF ordering within service function chains, then using a greedy heuristic with a carbon-aware cost function, and finally performing local search for load balancing refinement. The Slice Controller implements admission control and runtime scaling capabilities by monitoring SLA violations through exponentially weighted moving averages and triggering reactive reallocation whenever prediction error exceeds a defined threshold τ.

4 EXPERIMENTS

Carbon-Aware VNF Placement algorithm is described as follows.

Algorithm: CA-VNF-Placement(SFCs, nodes, λ_{pred}, CI_{pred}, t).

Input: Service function chains, available nodes, predicted traffic, predicted carbon intensity, time slot t.

Output: Placement decisions X, resource allocations R

1. Initialize: $X \leftarrow \varnothing, R \leftarrow 0$.

2. Compute priority scores for each slice k:

$$\pi_k \leftarrow \left(L_k^{\max} - D_k^{current}\right) / L_k^{\max} \quad \text{// Urgency metric.}$$

3. Sort slices by π_k in descending order $\rightarrow S_{sorted}$.

4. For each slice k in S_{sorted}:

a. For each VNF v in SFC_k:

I. Compute carbon-aware cost for each feasible node j:

$$cost_j \leftarrow \alpha \cdot CI_j(t) P_j(v) + \beta \cdot D_j(v) + \gamma \cdot c_j .$$

II. Select $j^* = argmincost_j$ subject to:

– $C_j - \sum_{i=1}^{n} R_{ij} \ge \rho_v$ (capacity),

– Queueing delay W_j satisfies SLA slack

III. Update: $x_{vj}^* \leftarrow 1, R_j^* \leftarrow R_j^* + \rho_v$,

b. If placement fails (infeasible):

– Attempt edge-to-core migration for delay-tolerant VNFs;

– If still infeasible, reject slice k (admission control).

5. Return X, R.

To handle traffic prediction errors and sudden bursts: reactive Scaling Trigger: If utilization $\rho_j(t) > \rho_{thresh}$ or queue length $Q_j(t) > Q^{\max}$: scale-out: activate standby VNF replicas, load redistribution: apply consistent hashing to migrate flows: proactive scaling based on prediction: at time t, for horizon $t+\Delta$: if $\chi_k(t+\Delta) > \lambda_k(t) + 2\sigma_k$, pre-warm resources, if $CI(t+\Delta) < CI(t) - \delta$, schedule deferrable workloads.

5 RESULTS

To assess the effectiveness of the proposed predictive carbon-aware orchestrator, we conduct a trace-driven simulation using a representative 5G/6G multi-layer infrastructure consisting of 10 RAN cells, four edge clusters (two servers each), and two core datacenters. Edge servers are provisioned with 64 CPU cores and 256 GB RAM, whereas core datacenters feature 256 CPU cores and 1 TB memory. The traffic model encompasses three representative slice types – eMBB, URLLC, and mMTC – capturing heterogeneous latency and compute requirements. Service time distributions follow realistic application behaviors, namely log-normal for eMBB, exponential for URLLC, and heavy-tailed Pareto for mMTC to reflect bursty IoT workloads. Historical carbon intensity traces from four European grid regions (2023–2024) are used to emulate spatio-temporal variations in carbon availability, updated at 15-minute intervals.

Four benchmark strategies serve as baselines (Table 1): (I) Latency-First (LF), which ignores carbon and prioritizes end-to-end delay; (II) Cost-Optimized (CO), which minimizes infrastructure cost without environmental considerations; (III) Carbon-Naive (CN), which applies average carbon intensity without predictive capabilities; and (IV) a hybrid state-of-the-art orchestration scheme from recent literature.

DOI 10.15588/1607-3274-2026-2-2

Table 1 – Comparison of orchestration strategies

Characteristic	Latency-First	Cost-Optimized	Carbon-Naïve	Hybrid orchestration
Carbon Awareness	No	No	Yes	Yes
Optimization Goal	End-to-end delay	Infrastructure cost	N/A	N/A
Predictive capabilities	No	No	No	Yes

Performance is evaluated using five classes of metrics: (I) carbon efficiency, defined as total CO_2 emissions over the evaluation interval; (II) cost efficiency in $/hour; (III) end-to-end latency, including both mean and 99th-percentile tail latency; (IV) SLA violation ratio; and (V) prediction accuracy measured by Mean Absolute Percentage Error (MAPE). These metrics jointly reflect environmental sustainability, economic feasibility, and QoS compliance.

6 DISCUSSION

The proposed framework achieves a 34% reduction in CO_2 emissions relative to the latency-first baseline and 28% improvement over the carbon-naive approach. This gain stems from both temporal shifting (aligning workload execution with predicted low-carbon periods) and spatial shifting (migrating workloads to greener geographical regions). By contrast, non-predictive schemes fail to exploit future carbon availability and therefore deliver inferior sustainability performance.

The orchestrator sustains high service quality while reducing emissions. eMBB slices incur only a modest increase in mean latency (45 ms vs. 42 ms for LF), while URLLC latency remains within strict QoS thresholds (3.2 ms, <5 ms SLA). mMTC traffic exhibits a 99th percentile latency of 180 ms, remaining below its 200 ms constraint. These results confirm that substantial carbon reductions are achievable without compromising SLA-sensitive services when optimization decisions are selectively applied to delay-tolerant slices.

Operational cost decreases by 22% relative to the commercial cost-optimized baseline, owing to three factors: reduced inter-layer traffic, improved workload placement efficiency, and natural correlation between low-carbon and off-peak pricing periods. Thus, carbon-aware optimization not only improves sustainability but also enhances economic viability.

The SLA violation rate remains below 1% across all slice types, demonstrating the robustness of the hybrid proactive – reactive scaling strategy. Violations occur primarily during rare traffic bursts where prediction error exceeds three standard deviations; however, real-time scaling mitigates most of these transient overloads.

Traffic forecasting yields MAPE values of 12.3% (1-hour horizon) and 18.7% (2-hour horizon), sufficient for proactive resource provisioning. Carbon intensity forecasting achieves 7.2% MAPE for one-hour ahead estimation, enabling accurate timing of workload migration decisions. These results confirm that even moderately accurate forecasting substantially improves orchestration outcomes.

We further investigate resilience to varying operational conditions. When traffic burstiness increases (coefficient of variation rising from 0.5 to 2.0), carbon savings reduce from 34% to 26% due to tighter queuing delays, yet remain significant. Under high-renewable energy availability (>60%), carbon savings amplify to 42%, indicating strong synergy between network orchestration and renewable penetration. The optimal prediction horizon is observed at 60–90 minutes, balancing forecast accuracy against orchestration benefits.

CONCLUSIONS

This paper presented a comprehensive framework for carbon-aware resource orchestration in 5G/6G networks that spans RAN, edge, and core domains. By combining rigorous queuing-theoretic modeling of bursty traffic with predictive carbon intensity forecasting, we achieve substantial environmental benefits-34% CO_2 reduction – while maintaining strict service quality guarantees. Our multi-objective optimization balances competing concerns of latency, cost, energy, and carbon emissions through adaptive weighting and dynamic resource scaling.

As mobile networks continue their exponential growth trajectory, integrating environmental considerations into core orchestration algorithms transitions from optional enhancement to operational imperative. Our work provides a theoretically grounded, empirically validated pathway toward sustainable next-generation wireless infrastructure.

The scientific novelty lies in the development of a predictive carbon-aware multi-layer slicing framework that jointly optimizes resource allocation across RAN, edge, and core domains under bursty traffic conditions. Unlike existing works, the proposed approach: integrates M/G/1 traffic modeling into slicing for realistic heavy-tailed fluctuations; combines traffic forecasting and carbon-intensity prediction into a unified orchestration pipeline; introduces a multi-objective optimization formulation that explicitly minimizes carbon emissions alongside latency, SLA compliance, and operational cost; demonstrates that predictive orchestration can simultaneously achieve environmental sustainability and QoS guarantees in next-generation mobile networks.

The practical significance of the proposed framework: reduce CO_2 emissions by up to 34% through time- and location-aware workload placement; lower operational expenditures by 22% via carbon-correlated cost-aware scheduling; maintain SLA violations below 1% even under bursty traffic; improve sustainability without the need for additional hardware deployment.

Prospects for further research include extending the framework to federated multi-operator slicing, integrating reinforcement learning-based online adaptation for long-horizon orchestration and implementing a prototype on

DOI 10.15588/1607-3274-2026-2-2

real testbeds for experimental validation beyond simulation.

ACKNOWLEDGEMENTS

The work is not supported by the state scientific research projects and is done by authors initiative.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Svitlana Sulima: the method of resource allocation; Ievgen Karashevych: experimental study of network orchestration method.

Data availability: The manuscript has no associated data.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Globa L., Sulima S., Skulysh M., Dovgyi S., Stryzhak O. Architecture and Operation Algorithms of Mobile Core Network with Virtualization. *IntechOpen,* 2020, 21 p. DOI: 10.5772/intechopen.89608
2. Yaqoob M., Trestian R., Tatipamula M., Nguyen H. X. Digital-Twin-Driven End-to-End Network Slicing Toward 6G. *IEEE Internet Computing*, 2024, Vol. 28, No. 2, pp. 47–55. DOI: 10.1109/MIC.2023.3332252
3. How 5G Network Slicing Expands Opportunities [Electronic resource]. *IPLOOK.* Available at: https://www.iplook.com/info/how-5g-network-slicing-expands-opportunities-i00232i1.html (accessed: 22.10.2025).
4. Tiwari K. Phulre A. K., Vishnu D. Enhancing Secure Key Management Techniques for Optimised 5G Network Slicing Security. *Applied Cybersecurity & Internet Governance,* 2024, Vol. 3, No. 2, pp. 170–210. DOI: 10.60097/ACIG/199725
5. Sargolzaei E., Rasti M., Khorsandi S. Topology and Energy Aware Approximate Algorithm for QoS-based Resource Slicing in 5G Core Networks. *IEEE Access,* 2025, Vol. 13, pp. 176885–176900. DOI: 10.1109/ACCESS.2025.3616851.
6. Khalili H., Papageorgiou A., Siddiqui S., Meixner C. C., Carrozzo G., Nejabati R., Simeonidou D. Network Slicing-aware NFV Orchestration for 5G Service Platforms. *Networks and Communications : European Conference EuCNC-2019, Valencia, 18–22 June, 2019 : proceedings*, pp. 25–30.
7. Kim Y., Lim H. Multi-Agent Reinforcement Learning-Based Resource Management for End-to-End Network Slicing. *IEEE Access,* 2021, Vol. 9, pp. 56178–56190.
8. Tsai C.-C., Lin F. J., Tanaka H. Evaluation of 5G Core Slicing on User Plane Function. *Communications and Network*, 2021, Vol. 13, No. 3, pp. 79–92. DOI: 10.4236/cn.2021.133007
9. Shao Y. Li R., Hu B., Wu Y., Zhao Z., Zhang H. Graph Attention Network-Based Multi-Agent Reinforcement Learning for Slicing Resource Management in Dense Cellular Network. *IEEE Transactions on Vehicular Technology*, 2021, Vol. 70, pp. 10792–10803.
10. Tariq M. A., Saad M. M., Ajmal M., Jeon D., Kim J., Kim D. Proactive Resource Management for Seamless Service: A Transition from 5G-Basic to 5G-Advanced Network Slicing. *Vehicular Technology : IEEE 100th Conference VTC2024-Fall. Washington DC, 7–10 October, 2024 : proceedings*, pp. 1–7. DOI: 10.1109/VTC2024-Fall63153.2024.10757954
11. Tselios C., Politis I., Amaxilatis D., Akrivopoulos O., Chatzigiannakis I., Panagiotakis S., Markakis E. K. Melding Fog Computing and IoT for Deploying Secure, Response-Capable Healthcare Services in 5G and Beyond. *Sensors,* 2022, Vol. 22, No. 9, Art. 3375, pp. 1–14. DOI: 10.3390/s22093375.
12. Salhab N., Langar R., Rahim R. 5G network slices resource orchestration using Machine Learning techniques. *Computer Networks,* 2021, Vol. 188, pp. 1–15. DOI: 10.1016/j.comnet.2021.107829.
13. Cai Y., Cheng P., Chen Z., Ding M., Vucetic B., Li Y. Deep Reinforcement Learning for Online Resource Allocation in Network Slicing. *IEEE Transactions on Mobile Computing,* 2024, Vol. 23, No. 6, pp. 7099–7116. DOI: 10.1109/TMC.2024.3344556.
14. Lin R., Liu H., Shan L., Zukerman M. Energy-aware Service Function Chaining Embedding in NFV Networks. *IEEE Transactions on Services Computing*, 2022, Vol. 99, pp. 1–14. DOI: 10.1109/TSC.2022.3162328

Received 23.10.2025.
Accepted 02.04.2026.
Published 26.06.2026.

DOI 10.15588/1607-3274-2026-2-2

УДК 004.9

ПРОГНОЗОВАНЕ БАГАТОРІВНЕВЕ РОЗДІЛЕННЯ РЕСУРСІВ ДЛЯ МЕРЕЖ 5G

Суліма С. В. – канд. техн. наук, доцент, доцент кафедри Інформаційних технологій в телекомунікаціях, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0000-0002-6333-7693.

Карашевич Є. Д. – аспірант кафедри Інформаційних технологій в телекомунікаціях, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0009-0000-0671-8185.

АНОТАЦІЯ

Актуальність. Швидке розгортання мереж 5G і поява архітектур 6G призводять до безпрецедентної гетерогенності трафіку і пачкоподібності в радіо-, периферійних і основних доменах. Тим часом, енергетичний слід мобільної інфраструктури стає основною проблемою сталого розвитку, оскільки викиди вуглецю все більше формують політику експлуатації мереж.

Мета роботи – розробка прогнозованої багаторівневої структури розподілу ресурсів з урахуванням вуглецевих викидів для мереж 5G/6G з периферійним ядром RAN, яка спільно оптимізує затримку, вартість, енергію та викиди вуглецю в умовах перевантаженого трафіку.

Метод. Запропонований підхід інтегрує модель черги на основі M/G/1 для точного представлення часу обслуговування «важких хвостів» і патернів прибуття; гібридне короткострокове/довгострокове прогнозування як транспортного навантаження, так і регіональної інтенсивності викидів вуглецю; і багатоцільову оптимізацію для розміщення VNF з урахуванням вуглецевих викидів і керування трафіком на різних рівнях мережі. Механізм проактивно-реактивної оркестровки виконує прогнозований попередній розподіл ресурсів і масштабування під час виконання.

Результати. Моделювання на основі трасування на репрезентативному багатошаровому тестовому стенді демонструє зниження викидів CO_2 на 34% порівняно з оркеструванням з урахуванням латентності, а також зниження операційних витрат на 22% і рівень порушень SLA <1%. Хвостова затримка залишається в межах порогових значень, визначених для кожного кадру, навіть під час пікових навантажень, підтверджуючи, що скорочення викидів вуглецю може бути досягнуто без погіршення якості обслуговування.

Висновки. Прогнозована оркестрація з урахуванням вуглецевих викидів у периферійних доменах RAN значно покращує екологічну та економічну ефективність, зберігаючи при цьому гарантії якості обслуговування. Результати підкреслюють важливість інтеграції оптимізації на основі прогнозування та реалістичного моделювання трафіку в архітектури розгалуження наступного покоління.

КЛЮЧОВІ СЛОВА: мережа з урахуванням вуглецевих викидів, багаторівневе розшарування, 5G/6G, прогнозна оркестровка, пачковий трафік, черги M/G/1, межа-хмара, стійка мережа.

ЛІТЕРАТУРА

1. Architecture and Operation Algorithms of Mobile Core Network with Virtualization / [L. Globa, S. Sulima, M. Skulysh, S. Dovgyi, O. Stryzhak]. – IntechOpen, 2020. – 21 p. DOI: 10.5772/intechopen.89608
2. Digital-Twin-Driven End-to-End Network Slicing Toward 6G / [M. Yaqoob, R. Trestian, M. Tatipamula, H. X. Nguyen] // IEEE Internet Computing. – 2024. – Vol. 28, No. 2. – P. 47–55. DOI: 10.1109/MIC.2023.3332252
3. How 5G Network Slicing Expands Opportunities [Електронний ресурс] // IPLOOK. – Режим доступу: https://www.iplook.com/info/how-5g-network-slicing-expands-opportunities-i00232i1.html (дата звернення: 22.10.2025). – Назва з екрану.
4. Tiwari K. Enhancing Secure Key Management Techniques for Optimised 5G Network Slicing Security / K. Tiwari, A. K. Phulre, D. Vishnu // Applied Cybersecurity & Internet Governance. – 2024. – Vol. 3, No. 2. – P. 170–210. DOI: 10.60097/ACIG/199725
5. Sargolzaei E. Topology and Energy Aware Approximate Algorithm for QoS-based Resource Slicing in 5G Core Networks / E. Sargolzaei, M. Rasti, S. Khorsandi // IEEE Access. – 2025. – Vol. 13. – P. 176885–176900. DOI: 10.1109/ACCESS.2025.3616851.
6. Network Slicing-aware NFV Orchestration for 5G Service Platforms / [H. Khalili, A. Papageorgiou, S. Siddiqui et al.] // Networks and Communications : European Conference EuCNC-2019, Valencia, 18–22 June, 2019 : proceedings. – P. 25–30.
7. Kim Y. Multi-Agent Reinforcement Learning-Based Resource Management for End-to-End Network Slicing / Y. Kim, H. Lim // IEEE Access. – 2021. – Vol. 9. – P. 56178–56190.
8. Tsai C.-C. Evaluation of 5G Core Slicing on User Plane Function / C.-C. Tsai, F. J. Lin, H. Tanaka // Communications and Network. – 2021. – Vol. 13, No. 3. – P. 79–92. DOI: 10.4236/cn.2021.133007
9. Graph Attention Network-Based Multi-Agent Reinforcement Learning for Slicing Resource Management in Dense Cellular Network / [Y. Shao, R. Li, B. Hu et al.] // IEEE Transactions on Vehicular Technology. – 2021. – Vol. 70. – P. 10792–10803.
10. Proactive Resource Management for Seamless Service: A Transition from 5G-Basic to 5G-Advanced Network Slicing / [M. A. Tariq, M. M. Saad, M. Ajmal et al.] // Vehicular Technology : IEEE 100th Conference VTC2024-Fall, Washington DC, 7–10 October, 2024 : proceedings. – P. 1–7. DOI: 10.1109/VTC2024-Fall63153.2024.10757954
11. Melding Fog Computing and IoT for Deploying Secure, Response-Capable Healthcare Services in 5G and Beyond / [C. Tselios, I. Politis, D. Amaxilatis et al.] // Sensors. – 2022. – Vol. 22, No. 9. – Art. 3375. – P. 1–14. DOI: 10.3390/s22093375.
12. Salhab N. 5G network slices resource orchestration using Machine Learning techniques / N. Salhab, R. Langar, R. Rahim. // Computer Networks. – 2021. – Vol. 188. – P. 1–15. DOI: 10.1016/j.comnet.2021.107829.
13. Deep Reinforcement Learning for Online Resource Allocation in Network Slicing / [Y. Cai, P. Cheng, Z. Chen et al.] // IEEE Transactions on Mobile Computing. – 2024. – Vol. 23, No. 6. – P. 7099–7116. DOI: 10.1109/TMC.2024.3344556.
14. Energy-aware Service Function Chaining Embedding in NFV Networks / [R. Lin, H. Liu, L. Shan, M. Zukerman] // IEEE Transactions on Services Computing. – 2022. – Vol. 99. – P. 1–14. DOI: 10.1109/TSC.2022.3162328

DOI 10.15588/1607-3274-2026-2-2