

Adaptive Spectrum-Aware Load Balancing for Multi-Access Edge Computing in Dense Vehicular Networks

Priya Ramanathan, *Member, IEEE*, Oskar Lindqvist, *Senior Member, IEEE*, and Naledi Mokoena, *Member, IEEE*

Abstract—The proliferation of connected and autonomous vehicles has placed unprecedented demand on multi-access edge computing (MEC) infrastructure deployed along dense urban corridors. Existing load-balancing schemes for vehicular MEC either assume static channel conditions or ignore the coupling between spectrum allocation and task-offloading latency, leading to congestion collapse at peak intersection density. This paper proposes a Spectrum-Aware Adaptive Load Balancer (SALB) that jointly optimizes edge-server task assignment and sub-channel allocation using a two-timescale Lyapunov drift-plus-penalty formulation. SALB decomposes the joint problem into a fast inner loop that solves per-slot channel assignment via a weighted bipartite matching relaxation, and a slow outer loop that adapts server admission thresholds based on a moving estimate of queue backlog. We evaluate SALB against three baselines – round-robin, greedy nearest-edge, and a deep reinforcement learning (DRL) scheme – on a simulated 5.8 GHz vehicular testbed spanning a 3.2 km urban corridor with 14 roadside units. SALB reduces mean task completion latency by 38.4% relative to greedy nearest-edge and by 21.7% relative to the DRL baseline at a vehicle density of 220 vehicles/km, while sustaining a 99.1% task admission rate under bursty arrival patterns. We further derive a closed-form bound on queue backlog under SALB and show empirically that the bound tracks simulated backlog within a 6% margin across all tested densities.

Index Terms—Multi-access edge computing, vehicular networks, load balancing, Lyapunov optimization, task offloading, spectrum allocation.

I. INTRODUCTION

THE shift toward cooperative perception and trajectory planning in connected vehicle fleets has moved a substantial share of onboard computation off the vehicle and onto roadside multi-access edge computing (MEC) nodes. Tasks such as cooperative object fusion, high-definition map alignment, and short-horizon trajectory re-planning are latency-critical: a completion deadline miss of even 40–60 ms can invalidate a maneuver decision at highway speed. As vehicle density increases at urban intersections and highway merge points, the roadside units (RSUs) that host MEC servers face two coupled bottlenecks – (i) contention for the limited number of orthogonal sub-channels available for task upload, and (ii) heterogeneous and time-varying compute queues at each edge server.

P. Ramanathan and N. Mokoena are with the Department of Electrical and Computer Engineering, Vertex Institute of Technology, Meridian City, MC 40221 (e-mail: p.ramanathan@vertex.edu; n.mokoena@vertex.edu).

O. Lindqvist is with the Synapse Research Laboratories, Apex Science Park, Nimbus, NB 71190 (e-mail: oskar.lindqvist@synapse-labs.org).

Manuscript received March 4, 2026; revised May 17, 2026.

Most prior work treats these two bottlenecks separately. Channel-allocation literature typically assumes a fixed, already-decided server assignment, while server load-balancing literature typically assumes an interference-free, infinite-capacity uplink. This separation is convenient for analysis but produces poor outcomes in practice: a load balancer that ignores channel contention may route many vehicles to a nominally under-loaded server that is, in fact, unreachable at usable rate because the serving RSU's sub-channels are saturated.

This paper makes the following contributions:

- 1) We formulate joint task-to-server assignment and sub-channel allocation for vehicular MEC as a two-timescale stochastic optimization problem and derive a Lyapunov drift-plus-penalty bound that admits a tractable per-slot decomposition.
- 2) We propose SALB, an online algorithm that solves the fast per-slot channel-assignment sub-problem via weighted bipartite matching and adapts slow-timescale admission thresholds from an exponentially weighted backlog estimate, requiring no offline training and no knowledge of future arrivals.
- 3) We validate SALB against round-robin, greedy nearest-edge, and a DRL baseline on a corridor-scale vehicular simulation, showing consistent latency and admission-rate improvements across a wide range of vehicle densities, and confirm the theoretical backlog bound empirically.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the system model. Section IV presents the SALB algorithm and its analysis. Section V reports experimental results, and Section VI concludes.

II. RELATED WORK

Task offloading in vehicular MEC has been studied extensively under the umbrella of vehicular fog and edge computing. Early work by Chen *et al.* [1] modeled offloading as a single-server queuing problem and derived stability conditions under Poisson arrivals, but did not consider multi-server contention. Subsequent work introduced multi-RSU cooperative offloading [2], using game-theoretic formulations to distribute load across neighboring servers; these approaches assume a shared, contention-free backhaul and do not model the wireless uplink explicitly.

On the spectrum-allocation side, sub-channel assignment for vehicle-to-infrastructure (V2I) links has typically been addressed via matching theory [3] or convex relaxation [4], optimizing throughput or fairness metrics independent of what the allocated capacity will be used for downstream. A smaller body of work has attempted joint optimization: Okafor and Reyes [5] proposed a joint admission-control and power-allocation scheme, but restricted server selection to a fixed nearest-RSU policy, which we show in Section V performs poorly under uneven spatial vehicle distribution.

Deep reinforcement learning has recently been applied to joint offloading and resource allocation [6], [7], and can, in principle, learn the coupling between channel state and queue backlog. However, DRL policies require substantial offline training on representative traffic traces and exhibit degraded performance under distribution shift – for instance, when an unmodeled event (a stalled vehicle, a temporary lane closure) produces a density profile absent from training. Our approach instead uses a model-based Lyapunov formulation that adapts online without retraining, at the cost of a mild sub-optimality gap relative to a fully clairvoyant policy, which we quantify in Section IV.

III. SYSTEM MODEL

We consider a corridor of M roadside units, indexed $m \in \{1, \dots, M\}$, each co-located with an MEC server of compute capacity C_m (in CPU cycles per second) and controlling K_m orthogonal sub-channels of bandwidth W Hz each. Time is slotted into intervals of duration τ ; within slot t , a set of vehicles $\mathcal{V}(t)$ generate offloadable tasks. Each task $i \in \mathcal{V}(t)$ is described by a tuple (d_i, ρ_i, δ_i) : input data size d_i (bits), required compute cycles per bit ρ_i , and a soft deadline δ_i (slots).

For a vehicle i served by RSU m over sub-channel k , the achievable uplink rate follows Shannon capacity under the measured signal-to-interference-plus-noise ratio $\gamma_{i,m,k}(t)$:

$$r_{i,m,k}(t) = W \log_2(1 + \gamma_{i,m,k}(t)). \quad (1)$$

The transmission delay for task i assigned to server m over channel k is $d_i/r_{i,m,k}(t)$, and the compute delay is $d_i\rho_i/C_m$ divided by the fraction of server capacity allotted to task i in that slot. Each server m maintains a queue backlog $Q_m(t)$ measured in remaining compute cycles. The backlog evolves as

$$Q_m(t+1) = \max[Q_m(t) - C_m\tau, 0] + \sum_{i \in \mathcal{A}_m(t)} d_i\rho_i, \quad (2)$$

where $\mathcal{A}_m(t) \subseteq \mathcal{V}(t)$ is the set of tasks admitted to server m in slot t . Our objective is to choose, for every slot t , an assignment $x_{i,m,k}(t) \in \{0, 1\}$ (task i to server m on channel k) to minimize the time-averaged expected task completion delay subject to (a) each vehicle assigned to at most one server-channel pair, (b) each channel used by at most one vehicle per slot, and (c) long-term queue stability, $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[Q_m(t)] < \infty$ for all m .

Algorithm 1 SALB – Spectrum-Aware Adaptive Load Balancer

```

1: Initialize  $Q_m(0) \leftarrow 0$  for all  $m$ ; set threshold  $\theta_m \leftarrow \theta_0$ .
2: for each slot  $t = 0, 1, 2, \dots$  do
3:   Measure  $\gamma_{i,m,k}(t)$  for all reachable  $(i, m, k)$  triples.
4:   Build weighted bipartite graph with edge weights from (3).
5:   Solve maximum-weight matching to obtain  $x_{i,m,k}(t)$ .
6:   Admit task  $i$  to server  $m$  only if  $Q_m(t) < \theta_m$ ; else defer to next slot if within deadline  $\delta_i$ , otherwise drop.
7:   Update  $Q_m(t+1)$  via (2).
8:   if  $t \bmod N = 0$  then
9:     Update  $\theta_m \leftarrow \theta_m + \eta (\bar{Q}_m^{\text{target}} - \hat{Q}_m(t))$ , where  $\hat{Q}_m(t)$  is an exponentially weighted moving average of  $Q_m$  over the last  $N$  slots.
10:  end if
11: end for

```

IV. SPECTRUM-AWARE ADAPTIVE LOAD BALANCING

A. Lyapunov Drift-Plus-Penalty Decomposition

Let $\mathbf{Q}(t) = (Q_1(t), \dots, Q_M(t))$ and define the quadratic Lyapunov function $L(\mathbf{Q}(t)) = \frac{1}{2} \sum_m Q_m(t)^2$. Standard drift-plus-penalty analysis [8] bounds the conditional drift $\Delta(t) = \mathbb{E}[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t)) \mid \mathbf{Q}(t)]$ plus a delay-penalty term $V \cdot \mathbb{E}[\text{delay}(t) \mid \mathbf{Q}(t)]$ by a constant B independent of the control policy, plus terms that are minimized by greedily solving, at every slot t :

$$\min_{x_{i,m,k}(t)} V \cdot \text{delay}(t) + \sum_m Q_m(t) \sum_{i,k} x_{i,m,k}(t) d_i \rho_i, \quad (3)$$

where $V > 0$ is a tunable parameter trading off delay against backlog stability. Because (3) decouples across channels given a fixed queue state $\mathbf{Q}(t)$, we solve it as a weighted bipartite matching between vehicles and (server, channel) pairs, with edge weight $V \cdot \tau / r_{i,m,k}(t) + Q_m(t) d_i \rho_i$.

B. Two-Timescale Algorithm

SALB separates the fast per-slot matching from a slow admission-threshold update, refreshed every N slots:

The per-slot matching step has worst-case complexity $O(|\mathcal{V}(t)|^2 \cdot \sum_m K_m)$ using the Hungarian algorithm restricted to reachable edges, which in our corridor topology (Section V) resolves in under 4 ms per slot on commodity hardware – well within a 20 ms slot duration. The threshold update requires only a scalar per server and executes at $1/N$ the frequency of the matching step, so its overhead is negligible.

C. Backlog Bound

Under mild boundedness assumptions on arrival rates and channel gains, the drift-plus-penalty analysis yields

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_m \mathbb{E}[Q_m(t)] \leq \frac{B + V \cdot \text{delay}^{\text{opt}}}{\epsilon}, \quad (4)$$

where ϵ is the slack between long-term arrival rate and aggregate server capacity, and $\text{delay}^{\text{opt}}$ is the delay achieved

TABLE I
MEAN TASK COMPLETION LATENCY (MS) AND ADMISSION RATE (%) AT
THREE VEHICLE DENSITIES

Scheme	80 veh/km	160 veh/km	220 veh/km
Round-Robin	61.2 / 99.8	118.7 / 92.4	201.5 / 74.6
Greedy Nearest-Edge	48.9 / 99.9	94.3 / 96.1	156.8 / 86.2
DRL	41.7 / 99.9	71.5 / 98.0	123.4 / 95.3
SALB (ours)	39.5 / 99.9	63.8 / 99.2	96.6 / 99.1

by an optimal clairvoyant policy with full future knowledge. Increasing V trades a $O(V)$ increase in worst-case backlog for delay approaching within $O(1/V)$ of delay^{opt}; we select V empirically in Section V by sweeping the delay-backlog trade-off curve.

V. EXPERIMENTAL RESULTS

A. Setup

We simulate a 3.2 km urban corridor with $M = 14$ RSUs spaced at irregular intervals reflecting realistic intersection geometry, each equipped with $K_m \in \{8, 12, 16\}$ sub-channels of $W = 180$ kHz (LTE-V2X resource-block granularity) and server capacity $C_m \in \{4, 6, 8\} \times 10^9$ cycles/s drawn from a heterogeneous mix consistent with staged RSU hardware rollouts. Vehicle arrivals follow a spatially non-uniform Poisson process calibrated to produce target densities of 40 to 260 vehicles/km, with task inter-arrival and size distributions drawn from a cooperative-perception workload trace (mean $d_i = 1.4$ Mb, mean $\rho_i = 620$ cycles/bit, deadline δ_i uniform over 3–8 slots at $\tau = 20$ ms). Each configuration is run for 5×10^4 slots across 20 random seeds; we report mean and 95% confidence intervals.

We compare SALB ($V = 850$, selected via the sweep in Fig. 2 of our extended technical report) against: (1) *Round-Robin*, which cycles task assignment across servers ignoring load or channel state; (2) *Greedy Nearest-Edge*, which always assigns a task to the geographically nearest RSU with capacity headroom; and (3) *DRL*, a deep deterministic policy gradient agent trained for 2×10^6 steps on a held-out set of density profiles drawn from the same generative process.

B. Latency and Admission Results

Table I reports mean task completion latency and admission rate at three representative densities. SALB consistently reduces latency relative to all baselines, with the margin widening as density increases – precisely the regime where channel contention and queue coupling matter most.

At 220 vehicles/km, SALB reduces mean latency by 38.4% relative to Greedy Nearest-Edge and 21.7% relative to DRL, while sustaining a 99.1% admission rate – 4.0 percentage points above DRL and 24.9 points above Greedy Nearest-Edge. Round-Robin’s admission rate collapses to 74.6% at high density because it continues to route tasks to congested servers regardless of backlog, producing widespread deadline misses.

C. Backlog Bound Validation

We compare the empirical time-averaged backlog against the closed-form bound of (4) across all tested densities. The bound tracks simulated backlog within a 6% margin at densities up to 220 vehicles/km, after which the bound-ness assumptions underlying the analysis begin to loosen as channel reachability becomes intermittent for vehicles at RSU cell edges. This confirms that V can be selected from the theoretical bound alone, without requiring an exhaustive empirical sweep in deployment settings where such a sweep is impractical.

D. Sensitivity to the Trade-off Parameter V

Sweeping V from 10^2 to 10^4 traces the expected delay-backlog trade-off curve. Latency improves monotonically with V up to approximately $V = 900$, beyond which marginal latency gains are outweighed by backlog growth that begins to threaten the soft deadlines of low-priority tasks. We therefore recommend $V \in [700, 950]$ for corridor deployments with task profiles similar to ours, and note that V should be re-tuned if the deadline distribution δ_i shifts substantially.

VI. CONCLUSION

We presented SALB, a two-timescale Lyapunov drift-plus-penalty algorithm for joint task-to-server assignment and sub-channel allocation in vehicular multi-access edge computing. By coupling a fast per-slot bipartite matching step with a slow admission-threshold adaptation, SALB avoids both the channel-blindness of pure load-balancing schemes and the training overhead and distribution-shift fragility of deep reinforcement learning approaches. Simulation on a corridor-scale vehicular testbed shows SALB reduces mean task completion latency by up to 38.4% relative to a strong greedy baseline and by 21.7% relative to a trained DRL policy at high vehicle density, while sustaining admission rates above 99%. Future work includes extending the model to multi-hop RSU cooperation, where a server at capacity can forward partially processed tasks to a neighboring RSU over a wired backhaul, and characterizing SALB’s robustness under adversarial jamming of the V2I control channel.

REFERENCES

- [1] L. Chen, H. Farrokhi, and M. Osei, “Queuing analysis of task offloading in vehicular fog computing under Poisson arrivals,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3781–3793, Apr. 2020.
- [2] S. Beaumont and T. Adeyemi, “Cooperative multi-RSU task offloading via coalition game theory,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5102–5114, Aug. 2021.
- [3] Y. Nakashima and R. Petrov, “Matching-theoretic sub-channel assignment for V2I uplink communications,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3990–4002, Jun. 2020.
- [4] D. Kowalski and F. Ibrahim, “Convex relaxation approaches to joint power and sub-channel allocation in vehicular networks,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, 2019, pp. 1–6.
- [5] C. Okafor and M. Reyes, “Joint admission control and power allocation for edge-assisted vehicular networks,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1654–1667, Feb. 2022.
- [6] J. Andersson, W. Tan, and P. Kallio, “Deep reinforcement learning for joint offloading and resource allocation in vehicular edge computing,” *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11256–11268, Jul. 2021.

- [7] A. Fontaine and S. Nkurunziza, "Actor-critic resource management for latency-constrained vehicular fog networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 19, no. 3, pp. 2540–2553, Sep. 2022.
- [8] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.
- [9] H. Vasquez and O. Lindqvist, "Delay-optimal resource allocation for dense small-cell vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7015–7028, Nov. 2020.
- [10] N. Mokoena, P. Ramanathan, and T. Halvorsen, "Backhaul-aware cooperative offloading for roadside edge servers," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

Priya Ramanathan received the B.Eng. degree in electrical engineering from the Vertex Institute of Technology, Meridian City, in 2016, and the Ph.D. degree in computer engineering from the same institution in 2021. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Vertex Institute of Technology. Her research interests include vehicular networking, edge computing resource allocation, and stochastic network optimization.

Oskar Lindqvist received the M.Sc. and Ph.D. degrees in telecommunications engineering from a consortium of Nordic technical universities in 2009 and 2013, respectively. Since 2015, he has been a Principal Researcher with Synapse Research Laboratories, Nimbus, where he leads the vehicular systems group. He has authored more than 60 papers on wireless resource allocation and holds four patents in adaptive spectrum management.

Naledi Mokoena received the B.Sc. degree in computer science from the University of Meridian in 2019 and is currently pursuing the Ph.D. degree in electrical and computer engineering at the Vertex Institute of Technology. Her research focuses on queueing-theoretic models for multi-access edge computing and their application to intelligent transportation systems.