

Spatiotemporal Analysis of Parallelized Computing at the Extreme Edge

Yasser Nabil, Graduate Student Member, IEEE, Mahmoud Abdelhadi, Sameh Sorour, Hesham ElSawy, Senior Member, IEEE, Sara A. Elsayed, Member, IEEE, and Hossam S. Hassanein, Fellow, IEEE

Abstract—Low-latency computational-task execution can be achieved by leveraging device-to-device offloading and parallel processing over nearby extreme edge devices (EEDs), a paradigm known as extreme edge computing (EEC). However, EEC performance is challenged by device spatial randomness with intermittent wireless connectivity, limited device computing power, time-varying availability, and device failures. This paper introduces a novel spatiotemporal analytical framework for EEC by integrating stochastic geometry with an absorbing continuous-time Markov chain (ACTMC) to capture the interplay between communication and computation. Modeling a large-scale millimeter-wave network, we derive tractable expressions for the average task response delay and the task completion probability under both random and location-aware EED selection. Numerical results quantify the impact of location-awareness and unveil the existence of an optimal task segmentation that minimizes delay, which depends on network parameters and EED capabilities. We also demonstrate that device failures and EED scarcity exacerbate delay, which can be mitigated through a collaborative load-balancing approach between EEC and Multi-Access Edge Computing (MEC) schemes. Simulations and sensitivity analyses validate the proposed framework and offer design insights for optimizing system performance.

Index Terms—Extreme edge computing (EEC), multi-access edge computing (MEC), task offloading, millimeter-wave communication, stochastic geometry, spatiotemporal modeling.

I. INTRODUCTION

Sixth-generation (6G) networks are anticipated to establish an infrastructure capable of supporting highly interconnected intelligent ecosystems [2], [3]. The anticipated 6G architecture features a diverse, intelligent, and perceptive structure facilitated by robust edge servers and distributed computing facilities [4], [5]. This enables a wide range of applications, such as digital twins, remote surgeries, smart cities, autonomous vehicles, industrial autonomy, and the Tactile Internet [2], [3]. Furthermore, 6G is projected to lead to an increase in device-to-device (D2D) connections, extensive utilization of artificial intelligence (AI), and a surge in the Internet of Things (IoT)

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant number: ALLRP 549919-20, and a grant from Distributive, Ltd.

Y. Nabil is with the Electrical and Computer Engineering Department, Queen's University, Kingston, Ontario, Canada. E-mail: yasser.nabil@queensu.ca.

M. Abdelhadi, S. Sorour, H. ElSawy, and H. S. Hassanein are with the School of Computing, Queen's University, Kingston, Ontario, Canada. E-mails: {m.abdelhadi, hesham.elsawy}@queensu.ca, and hossam@cs.queensu.ca.

S. A. Elsayed is with the Department of Computer Science, University of Calgary, Calgary, Alberta, Canada. E-mail: sara.elsayed@ucalgary.ca
Y. Nabil and M. Abdelhadi have contributed equally to this work.

This work was presented in part in [1].

services [2]–[5]. This is expected to trigger an unprecedented increase in data traffic and a corresponding need for extensive computations in the network.

Extensive computation is required for many AI workloads that underpin latency-sensitive IoT applications with strict quality-of-service (QoS) requirements [6], [7]. One solution to handle such extensive computations is to utilize cloud computing by offloading tasks to remote data centers. However, cloud computing fails to adequately satisfy latency-sensitive applications due to the distant geographical location of data centers and the huge traffic influx imposed at backhaul links [6], [7]. Multi-access edge computing (MEC) has emerged as a promising paradigm that can bring computing services closer to end devices, effectively reducing latency and meeting the increasing demands of IoT applications [8]. MEC platforms typically rely on edge servers co-located with or near base stations (BSs) to handle offloaded computational tasks, making efficient task offloading decisions crucial for achieving optimal MEC performance. The process of task offloading in MEC environments is significantly influenced by the availability, accessibility, and resilience of resources [9]. However, managing these factors can be costly and may not be applicable in certain scenarios. Additionally, the increasing number of devices utilizing MEC has given rise to the unresolved challenge of high congestion.

To overcome these limitations, extreme edge computing (EEC) offers an attractive alternative that leverages the abundant yet underutilized computational resources of IoT devices, referred to as extreme edge devices (EEDs), including smartphones, laptops, and connected vehicles [10]. While individual IoT devices possess limited processing power, their collective computational capabilities, when used in parallel, represent a significant untapped resource [11]. In EEC, these devices are harnessed to expand the computational resource pool, facilitate parallel processing, and improve task offloading by bringing the computing service much closer to end users, thus significantly reducing the response delay. Utilizing the abundant and underutilized computational resources of EEDs can also disrupt the dominance of traditional cloud service providers and network operators, fostering a more decentralized and democratized edge computing ecosystem with substantial advantages.

Despite its promising advantages, the EEC architecture faces several unique and interrelated challenges: (1) spatial randomness, (2) limited computational power of individual EEDs, (3) device vulnerability, and (4) temporal randomness. Spatial randomness arises from the highly dynamic network

topology, potentially leading to an insufficient number of EEDs in certain locations [12]. Unlike conventional MEC or cloud computing, EEDs have limited computational resources, making parallel task execution across multiple devices essential to meet performance requirements. Device vulnerability presents another significant challenge, as these user-owned devices are subject to intermittent availability, uncertainty, and higher failure risks, thereby requiring explicit reliability considerations [13]. Furthermore, temporal randomness emerges from fluctuations in offloading durations and task execution times, primarily due to the uncertainty associated with wireless channel conditions, signal-to-interference-plus-noise ratio (SINR) variability, task size diversity, and heterogeneous device capabilities. These intertwined challenges, including stochastic communication success and the temporal overlap between computation and communication, highlight the critical need for a rigorous spatiotemporal mathematical framework. Such a model is essential for accurately quantifying performance trade-offs in the EEC architecture, a gap that remains unaddressed in the existing literature.

In this paper, we quantify the interplay between communication and computation costs within large-scale millimeter-wave (mmWave) networks for EEC. Our primary contribution lies in developing the first spatiotemporal analysis for EEC, combining stochastic geometry (SG) and queueing theory, and uniquely employing an absorbing continuous-time Markov chain (ACTMC) to capture the dynamic interaction between task offloading via D2D communication and parallel computation across EEDs, which overlap in time. The proposed system partitions computational tasks into smaller segments, which are offloaded to multiple EEDs to accelerate execution.

The proposed spatiotemporal analysis is readily applicable to data-parallel IoT workloads. Representative examples include distributed machine learning inference on sensor readings or image batches and real-time video analytics [14], feature-map-partitioned deep-learning inference [15], and privacy-aware task segmentation [16]. More recently, such latency- and reliability-critical execution has become equally important for emerging edge-enabled digital-twin services. For instance, digital-twin pipelines may require the dynamic deployment and continuous update of human digital twins across edge servers to support timely task execution under mobility and time-varying conditions [17]. Likewise, interactive digital-twin services impose tight round-trip communication and computation constraints to deliver responsive feedback and user experience under evolving twin states [18]. In such digital-twin pipelines, many update and interaction workloads are naturally partitionable and can be opportunistically executed in parallel over nearby EEDs to meet stringent timeliness and reliability requirements, precisely the regime quantified by our spatiotemporal EEC analysis.

To this end, we analytically evaluate the average task response delay, a fundamental performance metric in EEC that reflects its viability for supporting latency-sensitive applications such as data-parallel distributed learning and real-time processing. Specifically, we utilize SG to derive the offloading success probability, accounting for device locations, mmWave antenna characteristics, channel conditions, and network-wide

interference. This probability determines the EED offloading rates utilized in our ACTMC model, which enables precise evaluation of the average task response delay. In addition to delay, we consider the task completion probability as a metric to evaluate system reliability, an essential consideration in failure-prone EEC environments. This metric captures the likelihood that all task segments are successfully executed. Together, these two metrics provide a meaningful assessment of EEC system performance and guide informed decisions on task segmentation levels, balancing both latency and reliability requirements.

To summarize, this paper makes the following contributions:

- We propose the first spatiotemporal analytical framework for EEC that integrates SG with an ACTMC to jointly model D2D offloading and parallel computation in large-scale networks.
- We derive tractable expressions for the average task response delay and task completion probability, characterizing the communication-computation trade-off that yields a delay-optimal task segmentation.
- We quantify the performance gains of location-aware EED selection and incorporate device failures, enabling explicit analysis of latency-reliability trade-off.
- We introduce a bias-based EEC-MEC collaboration scheme to mitigate system congestion in mmWave networks with limited line-of-sight (LoS) EED availability.
- We validate the analysis via Monte Carlo simulations and sensitivity studies, yielding robust design guidelines across diverse network conditions.

The remainder of the paper is organized as follows. Section II reviews the related work. Section III introduces the baseline spatiotemporal model with random EED selection under abundant EED availability. Section IV extends the model to practical scenarios with limited EED availability, incorporating location-aware selection, device failures, and EEC-MEC collaboration. Section V presents the numerical and simulation results, and Section VI concludes the paper and discusses future work.

II. RELATED WORK

Scheduling tasks in centralized cloud data centers has been extensively studied to improve the efficiency of parallel workflow execution across heterogeneous virtual machines (VMs). Complex applications are typically modeled as directed acyclic graphs (DAGs), where interdependent tasks are mapped to VMs and scheduled to optimize multiple objectives. Mohammadzadeh et al. investigated scientific workflow scheduling in green-cloud environments and proposed an improved chaotic binary Grey Wolf algorithm to minimize cost, makespan, and power consumption [19]. A related study broadens the optimization scope to a system-level multi-objective framework for enhanced resource efficiency and overall workflow performance in cloud data centers [20]. Li et al. [21] further extend workflow scheduling by incorporating security requirements, jointly optimizing service cost and data protection when deploying workflows across cloud-based VMs. Although effective for centralized infrastructures, these approaches are

not designed for the latency-sensitive and spatially distributed requirements of emerging 6G-enabled IoT and AI-driven applications, motivating the shift toward MEC.

MEC has therefore been widely investigated as a means to offload computation from resource-constrained devices to nearby edge servers. In [22], Jiang et al. formulate a real-time optimization framework for joint task offloading and resource allocation in MEC under a long-term energy constraint, where each task is treated as an indivisible unit executed either locally or at the edge. Liu et al. [23] develop an optimal stochastic computation offloading policy in a single-user MEC system, allowing tasks to be processed locally, at the MEC server, or in parallel across local and MEC processors. In [24], Chen et al. adopt a game-theoretic approach, designing an algorithm that converges to a Nash equilibrium and achieves performance and scalability when many devices share MEC resources.

Queueing-aware MEC control has been studied for delay-sensitive services; for example, Yi et al. jointly design multi-user computation offloading and uplink transmission scheduling with pricing-based incentives to mitigate congestion under random task arrivals [25]. Moreover, Cao and Cai [26] formulate multichannel-contention cloudlet offloading as a noncooperative game and propose a fully distributed learning algorithm that converges to a pure-strategy Nash equilibrium without information exchange among users. In more specialized scenarios, Moghaddasi et al. [27] study vehicular MEC and propose a double deep Q-network (DDQN)-based multi-objective offloading strategy that jointly optimizes latency, energy, and monetary cost while integrating a blockchain layer for data integrity and coordination. Furthermore, Rahmani et al. [28] focus on IoT-MEC energy management, developing a decentralized soft actor-critic framework for device-local, context-aware power control with limited MEC coordination and reporting substantial energy and battery-life gains in smart-home environments.

Beyond MEC, EEC further leverages nearby EEDs as cooperative workers for task execution. Azmy et al. develop an incentive-vacation queueing framework for reward-based EEC, where user-owned EEDs act as workers and incentive-coupled vacation queues with continuous-time Markov chains are used to characterize queueing delay and worker time in system, and to derive metrics for predicting worker participation dynamics under different incentive contracts [29]. Masoumi et al. apply EEC in industrial settings with mobile robots, using a hierarchical EEC/edge/cloud architecture with heuristic queue-aware scheduling and deadlock mitigation to coordinate movement and onboard processing tasks [30]. For digital twin services, El-Khatib et al. introduce a proactive scheme that maximizes a weighted service capacity objective by predicting EED resource usage and forming collaborating worker groups to execute partitioned subtasks under deadline constraints [11].

Learning-based orchestration has also been explored. Safavifar et al. propose a multi-objective deep reinforcement learning (DRL) workload orchestrator that assigns tasks to heterogeneous EEDs to reduce resource waste and energy consumption while maintaining a high task success rate [31]. Moreover, a DRL-based orchestrator for dependent composite

tasks in EEC has been proposed in [32], where applications are modeled as DAGs and decomposed into partitions that are offloaded to EEDs to minimize completion time and reduce MEC usage. From an experimental perspective, Drainakis et al. demonstrate service orchestration at the extreme edge over a 5G testbed, where an orchestrator manages containerized AI tasks on mobile and static EEDs via policy-based EED selection [33].

The works discussed above primarily focus on task offloading, resource allocation, and local dependability metrics under abstract network models, without explicitly capturing spatial randomness, network-wide interference, or the impact of node density and topology on performance. SG is therefore utilized, as in [12], [34], [35], to capture the effect of network geometry and interference in large-scale wireless systems with edge computing. In [12], task offloading in a mobile cloud computing network is analyzed under heterogeneous computational resources, and the network-wide outage probability is characterized. Elbayoumi et al. analyze edge computing in ultra-dense networks where small cells equipped with edge computing servers form a Poisson point process (PPP), and human-type users can associate with multiple small cells to partition and offload elastic tasks that are processed in parallel at the edge and locally [34]. More recently, Cheng et al. consider a MEC-aided uplink LoRa network with randomly distributed end devices modeled as a PPP and analyze the computation offloading success probability under interference and power control [35].

Recent efforts have combined queueing theory with SG to jointly capture network geometry and temporal dynamics, enabling a more complete spatiotemporal characterization of large-scale wireless systems [36], [37]. This spatiotemporal viewpoint has motivated several MEC studies that jointly account for communication and computation aspects of task execution. In [38], a spatiotemporal model is proposed for large MEC networks, where SG and queueing analysis are used to characterize both communication and computation latency. In [39], the scalability of MEC-enabled wireless networks is explored, and both communication and computation performance bounds are derived under a variety of network parameters. However, each task is modeled as an indivisible job, and a user can offload to at most a single MEC server. In [40], Gu et al. study a large-scale MEC wireless network in which tasks can be computed locally or offloaded to a MEC server, modeling the spatial distribution of access points (APs) and users via SG and employing a two-dimensional discrete-time Markov chain to capture the joint evolution of local-computation and offloading buffers and thereby characterize end-to-end task execution performance.

Other works have extended this spatiotemporal modeling to incorporate dependability and heterogeneous deployments. In [41], Emara et al. consider the joint impact of network interference and parallel computing with multiple VMs residing on the same edge server. Although computation is parallelized across VMs, each task is still offloaded to a single centralized MEC server. In [42], Park and Lee develop a spatiotemporal framework for MEC-enabled heterogeneous networks, where multi-tier MEC servers and users are modeled as PPPs and SG

is combined with an M/G/1 queueing model to characterize communication and computation latency. In [43], Gu et al. develop a spatiotemporal framework for MEC-enabled heterogeneous networks with communication-computation-aware user association, where multi-tier MEC APs and users are modeled as PPPs and SG is combined with queueing analysis to characterize the meta distribution of task offloading success and the resulting latency.

The above works are summarized in Table I, which highlights that existing spatiotemporal frameworks remain MEC-centric. In these models, tasks are typically executed at infrastructure MEC servers, and users cannot leverage nearby EEDs as cooperative workers within the EEC paradigm. To the best of our knowledge, this is the first work to develop a spatiotemporal analytical framework for EEC, jointly capturing the interplay between D2D connectivity, network-wide interference, and parallel computing across EEDs to analytically characterize EEC performance.

III. THE BASELINE SPATIOTEMPORAL ANALYSIS

This section presents a baseline spatiotemporal model, where EEDs are the only option that offers computational services. The EEDs are abundant, and their selection is made at random.

A. Baseline System Model

The computationally capable EEDs, also referred to as workers, are modeled via a PPP $\Phi \subset \mathbb{R}^2$ with intensity ν_w . The EEDs offer their computational services to resource-constrained devices (e.g., IoT), which hereafter are referred to as requesters. The requesters are spatially distributed according to an independent PPP $\Omega \subset \mathbb{R}^2$ with intensity ν_r . There is an edge orchestrator that can be a BS or an AP, which organizes the offloading process between workers and requesters. In particular, the EEDs that have available computational power register their availability at the edge orchestrator, which in turn informs each requester about the availability of proximate EEDs. Specifically, when a requester decides to offload a task to the surrounding EEDs, it requests the edge orchestrator to assign available nearby EED resources. In that context, the edge orchestrator does not have the location information, so it sends the devices in a random order. It is assumed in this model that $\nu_w \gg \nu_r$, and hence, the edge orchestrator can readily allocate a unique worker to each task segment without contention. To utilize parallel computing and reduce response delay, the requester divides each computational task into n smaller and equivalent segments to be offloaded and executed at different EEDs. Due to the heterogeneity of the computational powers of the EEDs, the execution time of each segment is exponentially distributed with mean $\frac{1}{n\mu_f}$, where μ_f is the task execution rate if computed at a single worker.

In compliance with 5G and beyond systems, the requesters utilize mmWave for D2D communications to offload segments to their proximate workers. The high vulnerability of mmWave communications to blockage is considered via the general LoS ball blockage model [44]. The devices within the distance of R_L from the requester are considered LoS devices, and

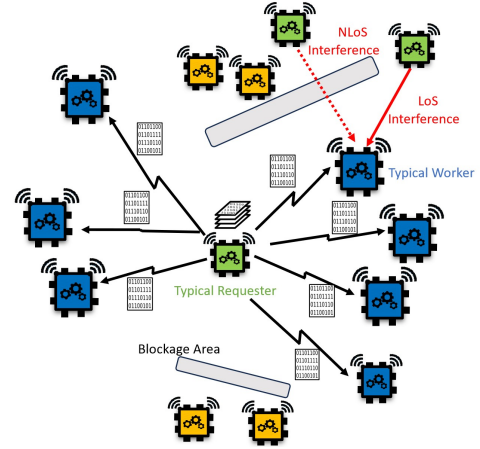


Fig. 1: The spatial system model: LoS workers (blue), NLoS workers (orange), and requesters (green). The typical requester offloads task segments to LoS workers. The typical worker receives the intended link from the typical requester, along with LoS and NLoS interference from other requesters.

otherwise, any device located beyond that point is considered a non-line of sight (NLoS) device. Distance-dependent power-law path-loss is considered with exponents α_L and α_N for LoS and NLoS devices, respectively. All transmissions experience Nakagami multipath fading. Hence, the channel power gains have independent and identical gamma distribution parameters N_L for LoS devices and N_N for NLoS devices.

Universal frequency reuse and constant transmit power are utilized by all requesters. The requester and workers deploy antenna arrays for mmWave beamforming. The widely adopted sectored antenna model approximates the array patterns [44]. Accordingly, the main lobe gain is M_x , the side lobe gain is m_x , and the 3-dB beamwidth is θ_x , where the subscript $x \in \{r, w\}$ is to differentiate between the antenna patterns of the requesters and workers. Without loss of generality, consider that a typical requester is located at the origin and can establish D2D links with proximate LoS EEDs only. Therefore, perfect antenna alignment is considered for the intended D2D link, and uniform random antenna alignment is considered for the interfering links. A pictorial illustration of the system model is shown in Fig. 1.

The requester is assumed to have one task divided into n independent and equal segments.¹ The segments are encapsulated into n packets transmitted via D2D communications to different proximate workers. The workers are sequentially allocated since a single mmWave interface is available at the requester. The workers are selected randomly from the list of available LoS EEDs provided by the edge orchestrator. Due to fluctuations in channel conditions, communication between the requester and the worker may encounter errors, which may require multiple attempts to deliver the segment and allocate the worker successfully. Each segment transmission attempt via D2D communication takes τ_c seconds. The worker

¹Tasks with inter-segment dependencies, intermediate synchronization, or precedence constraints, as in DAG-structured workflows, fall outside the present scope of our model.

TABLE I: Summary of edge computing related works

Work	MEC	EEC	Large-scale modeling (SG)	Spatiotemporal (SG+queueing)	Spatiotemporal EEC
MEC task offloading and control [22]–[28]	✓	×	×	×	×
EEC workload orchestration [11], [29]–[31]	×	✓	×	×	×
Hybrid MEC-EEC workload orchestration [32], [33]	✓	✓	×	×	×
SG-based works [12], [34], [35]	✓	×	✓	×	×
Spatiotemporal-based works [38]–[43]	✓	×	✓	✓	×
This work	✓	✓	✓	✓	✓

begins executing the segment immediately upon receiving the segment successfully. Upon receiving an acknowledgment (ACK) indicating a successful transmission, the requester offloads the remaining segments to other available LoS workers. Conversely, if a negative acknowledgment (NACK) is received, it indicates a transmission failure, prompting the requester to retransmit the same task segment until successful delivery. The ACK and NACK notifications are assumed to be transmitted over a perfect feedback channel. The result of each worker's assigned segment is returned to the requester as soon as that worker finishes executing it. We assume the communication time for returning these segment results is negligible, which aligns with our target domain of IoT monitoring and distributed inference pipelines. In such scenarios, outputs such as classification labels or control commands are compact relative to input data like sensor streams or high-resolution images. This simplification is also consistent with standard practice in related literature [22], [35], [43].

B. Offloading Success Probability with Random EED Selection

To calculate the average task response delay, we first need to obtain the average segment offloading time. The worker correctly receives the segment if the SINR is above a given threshold ξ . Otherwise, the segment has to be retransmitted. Hence, the first step in investigating the response delay is to find the D2D communication success probability between the requester and the randomly selected LoS worker. Such probability will be utilized later within an ACTMC to find the average task response delay. Following [44], the received SINR at the intended worker is given by

$$\text{SINR} = \frac{h_0 M_r M_w C_L r_0^{-\alpha_L}}{\sigma^2 + I_N + I_L}, \quad (1)$$

and the successful D2D transmission probability of a segment can be expressed as

$$p_s = \mathbb{P}\{\text{SINR} > \xi\} = \mathbb{P}\left\{\frac{h_0 M_r M_w C_L r_0^{-\alpha_L}}{\sigma^2 + I_N + I_L} > \xi\right\}. \quad (2)$$

Here, h_0 is the intended channel power gain, C_L is the intercept of the LoS channel, r_0 is the distance between the requester and the intended LoS worker, I_L is the aggregate interference from other active LoS requesters, I_N is the aggregate interference from other active NLoS requesters. Moreover, σ^2 denotes the normalized noise power, i.e., $\sigma^2 \triangleq \sigma_{\text{th}}^2 / P_t$, where σ_{th}^2 is the thermal noise power over bandwidth B and P_t is the transmit power of each active requester. Let $\Omega_L \subset \Omega$ and $\Omega_N = \Omega \setminus \{(\Omega_L) \cup (0, 0)\}$ be the point processes of the LoS

and NLoS requesters, respectively. Then, the LoS and NLoS interference terms as described in [44], are then expressed by

$$I_L = \sum_{i>0: \mathbf{x}_i \in \Omega_L} h_i D_i C_L \|\mathbf{x}_i\|^{-\alpha_L}, \quad (3)$$

and

$$I_N = \sum_{i>0: \mathbf{y}_i \in \Omega_N} g_i D_i C_N \|\mathbf{y}_i\|^{-\alpha_N}, \quad (4)$$

where h_i is the i^{th} LoS interfering link channel power gain, g_i is the i^{th} NLoS interfering link channel power gain, C_N is the intercept of the NLoS channel, $\|\cdot\|$ is the Euclidean norm, and D_i is the antenna gain for the i^{th} interfering requester in Ω_L or Ω_N . Given the sectorized antenna model and the uniformly random alignment between a typical worker and an interfering requester, D_i is a discrete random variable with four possible outcomes, each corresponding to a specific antenna gain scenario. These scenarios reflect the four possible alignments between the worker and the interferer requester, each with a specific probability. The distribution is $\mathbb{P}\{D_i = a_k\} = b_k$ for $k \in \{1, 2, 3, 4\}$, with a_k and b_k as defined in Table II.

The D2D transmission success probability given by (2) is characterized in Theorem 1.

Theorem 1: The spatially averaged probability of successful segment offloading via mmWave D2D communication to a randomly selected LoS worker from Φ_w is given by

$$p_s = \int_0^{R_L} \sum_{n=1}^{N_L} \binom{N_L}{n} \frac{2r_0(-1)^{n+1} e^{M_n(\xi)\sigma^2 - W_n(\xi) - Z_n(\xi)}}{R_L^2} dr_0, \quad (5)$$

where $M_n(\xi) = -\frac{\eta_L n r_0^{\alpha_L} \xi}{C_L M_r M_w}$, while $W_n(\xi)$ and $Z_n(\xi)$ are given by

$$W_n(\xi) = 2\pi\nu_r \sum_{k=1}^4 b_k \int_0^{R_L} \left(1 - \frac{1}{\left(1 + \frac{\eta_L \bar{a}_k n \xi \left(\frac{r_0}{x}\right)^{\alpha_L}}{N_L}\right)^{N_L}}\right) x dx, \quad (6)$$

$$Z_n(\xi) = 2\pi\nu_r \sum_{k=1}^4 b_k \int_{R_L}^{\infty} \left(1 - \frac{1}{\left(1 + \frac{n_L \bar{a}_k n \xi C_N r_0^{\alpha_L}}{C_L x^{\alpha_N} N_N}\right)^{N_N}}\right) x dx. \quad (7)$$

Here, $\bar{a}_k = \frac{a_k}{M_r M_w}$, and b_k along with a_k for $1 \leq k \leq 4$ are defined in Table II.

Proof: The proof can be found in Appendix A ■

C. Average Task Response Delay Calculation

The task response delay is defined as the time needed to process the n segments, beginning when the requester starts offloading the first segment to the allocated EED

TABLE II: Directivity Gain and Probability

k	a_k	b_k
1	$M_w M_r$	$\frac{\theta_w}{2\pi} \frac{\theta_r}{2\pi}$
2	$M_w m_r$	$\frac{\theta_w}{2\pi} \left(1 - \frac{\theta_r}{2\pi}\right)$
3	$m_w M_r$	$\left(1 - \frac{\theta_w}{2\pi}\right) \frac{\theta_r}{2\pi}$
4	$m_w m_r$	$\left(1 - \frac{\theta_w}{2\pi}\right) \left(1 - \frac{\theta_r}{2\pi}\right)$

and concluding when all n segments have been executed. This delay encompasses both communication and computation components, along with their interactions. However, due to the randomness in factors such as EED locations and availability, channel gains, computational power, and failure rates, the delay calculated in this study is represented as an average measure, referred to as the average task response delay. Note that the average time that one segment takes to be executed is $T_o = \tau_h + \tau_f$, where τ_h is the average offloading time that the requester takes to offload a segment to a randomly selected EED, and τ_f is the average time the segment takes to be executed at the intended EED. In this context, $\tau_h = \tau_c/p_s$, where p_s is the probability given in Theorem 1, and τ_c is the average D2D communication time in mmWave networks.

The average task response delay cannot be simply represented as the sum of individual segment delays ($\neq nT_o$), as this would ignore both the parallel processing within the system and the overlap between communication and computation times. The system's complexity, influenced by the interactions between simultaneous segment processing, the stochastic nature of communication offloading, and the overlapping of communication and computation times, combined with the fact that allocation and completion events can happen at any moment, requires modeling using an ACTMC. To this end, the successful offloading probability estimated in Theorem 1 is a core building block of the ACTMC, where the average offloading rate $\lambda_h = 1/\tau_h$. Next, we delve into the foundational ACTMC and embedded discrete-time Markov chain (EDTMC) employed.

1) *ACTMC and EDTMC*: The state set of the ACTMC is represented as $\mathcal{S} = \{\mathbf{z} = (x_f, x_c) \mid \sum_j x_j \leq n; j \in \{f, c\}\}$, where $x_f \in \{0, 1, 2, \dots, n\}$ denotes the number of workers that have finished their assigned segments successfully, and $x_c \in \{0, 1, 2, \dots, n\}$ denotes the number of workers that are executing the assigned segments. For each task, ACTMC starts at the state $\mathbf{z}_1 = (0, 0)$, where the requester has a task that is sliced to n segments but has not yet allocated any worker. Each time the requester succeeds in allocating a LoS EED via mmWave D2D transmission, a transition occurs from the current state $\mathbf{z}_i = (x_f, x_c)$ to the next state $\mathbf{z}_j = (x_f, x_c + 1)$. Moreover, each time a worker is retired because of segment completion, a transition from state $\mathbf{z}_i = (x_f, x_c)$ to $\mathbf{z}_j = (x_f + 1, x_c - 1)$ occurs. Since the requester needs only n workers, then $x_c + x_f \leq n$ and $\mathbf{z}_L = (n, 0)$ is the absorbing state that implies the termination of the ACTMC, where L is the total number of states in the system.

Following the criterion mentioned above, segments offloading and execution at the EEDs can be tracked with an ACTMC

with the following two-level hierarchical generator matrix

$$\mathbf{Q} = \begin{matrix} x_f & 0 & 1 & 2 & 3 & \dots & n \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{matrix} & \begin{pmatrix} \mathbf{K}_0 & \mathbf{H}_{0,1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_1 & \mathbf{H}_{1,2} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_2 & \mathbf{H}_{2,3} & \ddots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{K}_{n-1} & \mathbf{H}_{n-1,n} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{matrix}, \quad (8)$$

where \mathbf{Q} is a block matrix of size $(n+1) \times (n+1)$ that tracks the number of finished workers x_f . Since the task is finished upon the completion of the n segments, then the state $x_f = n$ is the absorbing state that indicates the termination of the edge computing. Within each level of \mathbf{Q} , the sub-matrices \mathbf{K}_m and $\mathbf{H}_{m,m+1}$ track the number of allocated workers x_c .² Exploiting the fact that $x_c + x_f \leq n$, the matrix $\mathbf{H}_{m,m+1}$ is of size $(n-m) \times (n-m-1)$ that tracks x_c due to the completion of a segment by any of the workers. Let $\mathbf{H}_{m,m+1}(i, j)$, with $i \in \{0, 1, 2, \dots, n-m\}$ and $j \in \{0, 1, 2, \dots, n-m-1\}$, denote the (i, j) th element of the matrix $\mathbf{H}_{m,m+1}$. Then, due to the parallelism in the computing at the EEDs along with the fact that only one worker can finish at a given instance, $\mathbf{H}_{m,m+1}$ is given by

$$\mathbf{H}_{m,m+1}(i, j) = \begin{cases} i\mu_f, & i = j + 1, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Using a similar argument, the matrix \mathbf{K}_m is of size $(n-m+1) \times (n-m+1)$ that tracks x_c upon allocating new workers. Let $\mathbf{K}_m(i, j)$, with $i, j \in \{0, 1, 2, \dots, n-m\}$ denote the (i, j) th element of the matrix \mathbf{K}_m . Accordingly, due to the sequential worker allocation, $\mathbf{K}_m(i, j)$ is given by

$$\mathbf{K}_m(i, j) = \begin{cases} -(\lambda_h + i\mu_f), & i = j \text{ and } i < n - m, \\ \lambda_h, & i = j - 1 \text{ and } i < n - m, \\ -(n - m)\mu_f, & i = j = n - m, \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where $\lambda_h = p_s/\tau_c$ is the offloading rate, p_s is the D2D transmission success probability given in (5), and τ_c is the time required for each D2D transmission attempt.

The average task response delay cannot be directly obtained for the matrix \mathbf{Q} . Instead, we first need to obtain the EDTMC of \mathbf{Q} and the average sojourn time at each state. The EDTMC

²The hierarchical structure of the proposed ACTMC enables scalable analysis and eliminates the need to visualize the full $(n+1) \times (n+1)$ generator matrix, whose entries are submatrices.

of \mathbf{Q} is given by

$$\mathbf{P} = \begin{matrix} x_f & 0 & 1 & 2 & 3 & \dots & n \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{matrix} & \begin{pmatrix} \mathcal{K}_0 & \mathcal{H}_{0,1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathcal{K}_1 & \mathcal{H}_{1,2} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{K}_2 & \mathcal{H}_{2,3} & \ddots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathcal{K}_{n-1} & \mathcal{H}_{n-1,n} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} \end{matrix}, \quad (11)$$

where \mathcal{K}_m and $\mathcal{H}_{m,m+1}$ track the transition probabilities due to worker allocation and segment completion, respectively. Moreover, the matrices \mathcal{K}_m and $\mathcal{H}_{m,m+1}$ are given by

$$\mathcal{K}_m(i, j) = \begin{cases} \frac{\lambda_h}{\lambda_h + i\mu_f}, & i = j - 1, \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

$$\mathcal{H}_{m,m+1}(i, j) = \begin{cases} \frac{i\mu_f}{\lambda_h + i\mu_f}, & i = j + 1 \text{ and } i < n - m, \\ 1, & i = n - m, j = n - m - 1, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

2) *Average Time until Absorption:* After formulating the ACTMC and obtaining the matrices \mathbf{Q} and \mathbf{P} , we now utilize those to calculate the average task response delay. The ACTMC has an absorbing state that is reached once all n segments have been successfully executed at the allocated EEDs. Based on that, the average task response delay is equivalent to the average time until absorption in that state. To calculate the average time until absorption, let $x_{c_i} \in \mathbf{z}_i$ be the number of allocated workers in state \mathbf{z}_i , then the average sojourn time $t_{\mathbf{z}_i, \mathbf{z}_j}$ is given by

$$t_{\mathbf{z}_i, \mathbf{z}_j} = \begin{cases} \frac{1}{x_{c_i}\mu_f}, & \text{if the transition from } \mathbf{z}_i \text{ to } \mathbf{z}_j \\ & \text{is due to segment completion,} \\ \frac{1}{\lambda_h}, & \text{if the transition from } \mathbf{z}_i \text{ to } \mathbf{z}_j \\ & \text{is due to worker allocation.} \end{cases} \quad (14)$$

Equipped with \mathbf{P} and $t_{\mathbf{z}_i, \mathbf{z}_j}$, the average task response delay is given in Theorem 2.

Theorem 2: The average task response delay in the extreme edge computing networks with mmWave D2D communications and n randomly allocated workers is given by

$$T_A = \alpha(\mathbf{I} - \mathbf{P}_T)^{-1}\mathbf{w}, \quad (15)$$

where $\alpha = [1, 0, 0, \dots, 0]$ with a dimension of $1 \times L$ represents the system's initial state, \mathbf{I} is the identity matrix, \mathbf{P}_T is the transition probability of the transient states only in \mathbf{P} , which is obtained by excluding the transitions to the absorbing state (the last row and column of \mathbf{P}). The column vector \mathbf{w} contains the average sojourn times at states \mathbf{z}_i , which are given by $w_{\mathbf{z}_i} = \sum_{\mathbf{z}_j} \mathbf{P}(\mathbf{z}_i, \mathbf{z}_j)t_{\mathbf{z}_i, \mathbf{z}_j}$, where $\mathbf{P}(\mathbf{z}_i, \mathbf{z}_j)$ is the transition probability from state \mathbf{z}_i to \mathbf{z}_j .³

Proof: The proof can be found in Appendix B ■

³In line with the hierarchical structure of \mathbf{P} , we use two-dimensional indexing for its elements. Specifically, $\mathbf{P}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{P}((x_{f_i}, x_{c_i}), (x_{f_j}, x_{c_j}))$ is the (x_{c_i}, x_{c_j}) element of the (x_{f_i}, x_{f_j}) sub-matrix in \mathbf{P} .

IV. ADVANCED SPATIOTEMPORAL ANALYSIS

To address the limitations of the baseline model, we introduce an advanced model that analyzes location-aware EED selection and accounts for potential EED failures. In addition, we propose task completion probability as a reliability metric, particularly relevant in scenarios with limited and/or failure-prone workers. This metric enables the quantification of system robustness under uncertainty and further enriches the performance evaluation of EEC environments by incorporating reliability in addition to latency. Furthermore, we investigate the impact of limited EED intensity relative to requester intensity on EEC performance and examine a bias model that enables collaboration between EEC and MEC, aiming to enhance system performance, particularly when contention over available LoS EEDs is high.

A. Advanced System Model

The advanced system model still shares some similarities with the baseline model described in Section III-A. Such similarities include the fact that workers and requesters are still modeled as PPPs with intensities ν_w and ν_r , respectively. In addition, the requester can only allocate the LoS EEDs due to blockages. Thus, offloading is performed after obtaining information on surrounding LoS EEDs from the edge orchestrator, which maintains EED availability information. Unlike the baseline model that assumes no scarcity of available EED, the advanced system model considers a limited number of available EEDs for each requester. In this case, when a requester decides to offload a task, the orchestrator maintains a pool with a limited number of EEDs, where the average number of available EEDs is $\nu_w \pi R_L^2$.

Moreover, when a requester probes the orchestrator for information about surrounding LoS EEDs, their locations are also included in the provided EED information. As a result, the offloading shifts from randomly selecting a device to preferring the closest i^{th} device for the i^{th} offloading action. This refined approach aims to improve the probability of successful offloading. By selecting a nearby device, the signal quality improves, leading to a decrease in path loss and an overall increase in both the successful offloading probability and the offloading rate.

In addition, effectively handling EED failures is essential for enabling realistic EEC operations; therefore, failure events are explicitly considered. Specifically, if an EED fails during task execution, the requester allocates a replacement EED. It is important to note that the execution time impacts the failure likelihood: the longer a device operates to complete a task, the more exposed it is to disruptions and dropouts. To quantify this, we define the failure rate as $\gamma = \frac{\mu_f}{l}$, where l represents the system reliability parameter, indicating that an EED, on average, fails l times less frequently than it successfully executes a task. For a task divided into n segments, the failure rate for each device is expressed as $\gamma_n = \frac{\gamma}{n}$.

Finally, to address congestion in practical scenarios resulting from multiple requesters competing for the limited available LoS EEDs, a collaborative offloading approach involving both

EEC and MEC is explored to reduce the average response delay. This method considers a bias factor, denoted by α , which represents the proportion of requesters offloading their tasks to EEDs.

B. Distance-based Successful Offloading Probability

Since EED allocation is done by selecting the closest device to the requester. Let $\mathbf{R} = \{R_{(1)}, R_{(2)}, \dots, R_{(k)}, \dots, R_{(n)}\}$ be the sorted distance vector of all the LoS EEDs, where k represents the rank of the EED in the sorted vector, such that $R_{(1)} = \min\{\mathbf{R}\}$ and $R_{(n)} = \max\{\mathbf{R}\}$. Theorem 3 represents the segment's successful offloading probability when selecting a new device based on its rank.

Theorem 3: The spatially averaged probability of successful segment offloading via mmWave D2D communications for a distance-based selected LoS worker is given by

$$p_{s_{(k)}} = \int_0^{R_L} \sum_{n=1}^{N_L} \binom{N_L}{n} (-1)^{n+1} e^{M_n(\xi)\sigma^2 - W_n(\xi) - Z_n(\xi)} f_{(k)}(r_0) dr_0, \quad (16)$$

where $M_n(\xi) = -\frac{\eta_L n r_0^{\alpha} \xi}{C_L M_r M_n}$, $W_n(\xi)$ and $Z_n(\xi)$ are given in (6) and (7), $f(x) = 2r_0/R_L^2$, $F(x) = r_0^2/R_L^2$, $V = \pi\nu_w R_L^2$, and $f_{(k)}(x)$ is given by

$$f_{(k)}(x) = \frac{V^k e^{-V} f(x) F(x)^{k-1}}{(k-1)!} e^{-V[F(x)-1]}. \quad (17)$$

Proof: The proof can be found in Appendix C ■

The offloading probability $p_{s_{(k)}}$ requires changing the ACTMC and EDTMC to be level-dependent, which means that any selected device has its own offloading rate. This offloading rate is represented as $\lambda_{h_k} = p_{s_k}/\tau_c$, where p_{s_k} is the successful offloading probability of the k^{th} closest EED to the requester. The updated matrices are given as follows:

$$\mathbf{K}_m(i, j) = \begin{cases} -(\lambda_{h_{i+1}} + i\mu_f), & i = j \text{ and } i < n - m, \\ \lambda_{h_{i+1}}, & i = j - 1 \text{ and } i < n - m, \\ -(n - m)\mu_f, & i = j = n - m, \\ 0, & \text{otherwise,} \end{cases} \quad (18)$$

$$\mathcal{K}_m(i, j) = \begin{cases} \frac{\lambda_{h_{i+1}}}{\lambda_{h_{i+1}} + i\mu_f}, & i = j - 1, \\ 0, & \text{otherwise,} \end{cases} \quad (19)$$

and

$$\mathcal{H}_{m,m+1}(i, j) = \begin{cases} \frac{i\mu_f}{\lambda_{h_{i+1}} + i\mu_f}, & i = j + 1 \text{ and } i < n - m, \\ 1, & i = n - m, j = n - m - 1, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

The practical implementation of distance-based EED selection relies on efficient location acquisition mechanisms with minimal overhead. The edge orchestrator can obtain EED location information through several methods native to modern wireless systems. The orchestrator can leverage existing uplink reference signals, already transmitted by EEDs for standard communication functions like channel state information (CSI) estimation, to determine EED locations [45]. This approach requires no dedicated additional signaling with

the EEDs, shifting the computational burden entirely to the orchestrator [45]. Alternatively, EEDs can self-report their positions using onboard GNSS/GPS capabilities by appending location data to the registration messages they send to the orchestrator to indicate their availability [46]. This method introduces only a few bytes of overhead and operates at a low frequency relative to task offloading cycles, ensuring the signaling cost remains minimal [46]. Looking forward, emerging 6G Integrated Sensing and Communication (ISAC) paradigms are expected to further streamline this process, as high-precision device location can be inferred directly by analyzing communication signals and their interactions with the environment, thereby eliminating dedicated positioning signaling and its associated overhead [47].

C. Modeling EEDs Failure

To reflect the changes in the proposed model after introducing the system reliability parameter, the matrices K_m and H_m are now represented as follows:

$$K_m(i, j) = \begin{cases} i\gamma_n, & j = i - 1, \\ -i(\gamma_n + \mu_f) - \lambda_{h_{i+1}}, & i = j, i < n - m, \\ -(n - m)(\gamma_n + \mu_f), & i = j = n - m, \\ \lambda_{h_{i+1}}, & j = i + 1, i < n - m, \\ 0, & \text{otherwise,} \end{cases} \quad (21)$$

$$H_m(i, j) = \begin{cases} i\mu_f, & j = i - 1, \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

Consequently, $\mathcal{K}_m(i, j)$ and $\mathcal{H}_{m,m+1}(i, j)$ are modified as follows:

$$\mathcal{K}_m(i, j) = \begin{cases} \frac{i\gamma_n}{i(\gamma_n + \mu_f) + \lambda_{h_{i+1}}}, & j = i - 1, i < n - m, \\ \frac{\gamma_n}{\gamma_n + \mu_f}, & i = n - m, j = n - m - 1, \\ \frac{\lambda_{h_{i+1}}}{i(\gamma_n + \mu_f) + \lambda_{h_{i+1}}}, & j = i + 1, i < n - m, \\ 0, & \text{otherwise,} \end{cases} \quad (23)$$

$$\mathcal{H}_{m,m+1}(i, j) = \begin{cases} \frac{i\mu_f}{i(\gamma_n + \mu_f) + \lambda_{h_{i+1}}}, & j = i - 1, i < n - m, \\ \frac{\mu_f}{\gamma_n + \mu_f}, & i = n - m, j = n - m - 1, \\ 0, & \text{otherwise.} \end{cases} \quad (24)$$

Moreover, the average sojourn time $t_{z_i, z_j} = \frac{1}{x_{ci}\gamma_n}$ if the transition from z_i to z_j is due to a failure.

D. Task Completion Probability

To further assess the resilience of the EEC system under realistic conditions, we introduce the task completion probability metric. It quantifies the likelihood that all parallelized task segments are successfully executed by dynamically recruited EEDs under conditions of limited worker availability and/or potential device failures. While the average task response delay serves as the primary metric for evaluating latency performance and identifying the optimal number of task segments n , it does not guarantee successful task completion. In contrast, task completion probability offers insight into the reliability of

the system by capturing the probability that all task segments are completed successfully. By analyzing both metrics jointly, we can determine the segmentation level n that satisfies not only delay minimization but also a desired reliability threshold. This enables informed decision-making for applications with varying priorities. For instance, safety-critical systems may prioritize reliability, whereas latency-sensitive tasks, such as real-time video processing, may focus on minimizing delay.

Mathematically, let $\rho_t(i)$ denote the probability that all n task segments are successfully completed, starting from system state i . This probability is computed recursively as:

$$\rho_t(i) = \begin{cases} 1, & \text{if } i \text{ is a success state,} \\ 0, & \text{if } i \text{ is a failure state,} \\ \sum_j P(i, j) \cdot \rho_t(j), & \text{otherwise (transient state),} \end{cases} \quad (25)$$

where $P(i, j)$ is the transition probability from state i to state j , and j indexes the successor state. A *success state* occurs when all n segments have been executed, whereas a *failure state* corresponds to scenarios in which the system exhausts its capacity to recover due to persistent failures and/or low worker intensity. The system-wide task completion probability, denoted by ρ_t , is defined as $\rho_t(0)$, where $i = 0$ corresponds to the initial state with no workers recruited and no segments completed. The transition probabilities $P(i, j)$ are derived from the failure-aware EDTMC, capturing the effects of failure events and recovery dynamics.

E. Worker Status and EED Bias Factor

Let $\alpha \in [0, 1]$ denote the bias factor, representing the fraction of requesters that offload to EEDs. This yields a requester intensity of $\nu_{r,\alpha} = \alpha\nu_r$ for EEC offloading, while the remaining $(1 - \alpha)\nu_r$ offload to MEC. This approach enables studying a combined EEC-MEC offloading strategy that balances computational load between both resources.

The availability of a worker (EED) depends on the following parameters:

- 1) The intensity of other workers ν_w , since the probability of an EED being available decreases with fewer workers.
- 2) The task execution rate μ_f , since the higher the execution rate, the higher the probability that the EED will be idle.
- 3) The intensity of the requesters ν_r , since each requester needs to allocate EEDs to execute its task, and thus the higher the number of requesters, the higher the probability that the EED will be busy executing a task.

Consequently, the status of each worker is represented by a continuous-time Markov chain (CTMC), where the worker can be in one of two states: *idle*, indicating it has no current task segment and is ready to receive one, and *busy*, indicating it is actively computing a segment. Fig. 2 illustrates the states and transitions in the worker's CTMC.

To model congestion, this CTMC determines the intensity of idle EEDs in scenarios where requesters compete for available EEDs. Let $\pi = \{\pi_{idle}, \pi_{busy}\}$ be the vector that represents the probability that an EED can be at the *idle* state or the *busy* state, respectively. The value of π is obtained by solving

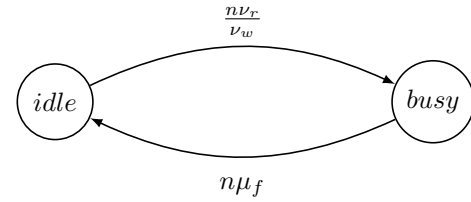


Fig. 2: Worker CTMC

$\pi\mathbf{Q} = \mathbf{0}$ and $\sum_s \pi_s = 1$ where $s \in \{idle, busy\}$ and \mathbf{Q} is the state transition matrix of the CTMC. After solving, the value of π_{idle} , which reflects the probability of an EED being idle, is given by

$$\pi_{idle} = \frac{\mu_f}{\mu_f + \frac{\nu_r}{\nu_w}}. \quad (26)$$

Note that the steady state solution π_{idle} does not depend on the number of task segments n . This is because, as the number of task segments increases, the probability of an EED being assigned a task and becoming busy also increases. However, the EED will complete its task more rapidly, returning to an idle state again.

Utilizing π_{idle} , the intensity of the EEDs that are idle is $\nu_{w_{idle}} = \pi_{idle} * \nu_w$, and the intensity of the EEDs that are busy is $\nu_{w_{busy}} = \nu_w - \nu_{w_{idle}}$. As explained before, the value of $\nu_{w_{idle}}$ depends on the value of ν_r , so low values of the bias factor α can be utilized to reduce the intensity of the requesters that will offload to EEDs, which will increase the number of available EEDs. Conversely, decreasing the value of α will increase the load on the MEC. To achieve the balance in the average response delay in both EEDs and MEC, let $\tau_\alpha = \alpha * \tau_{EEDs} + (1 - \alpha) * \tau_{MEC}$ be the average response delay for the EEDs and the MEC, and the optimal value of α will be the one that results in the lowest value of τ_α .

It is worth highlighting that by congestion we mean scenarios where requesters face difficulty in finding available EEDs due to system wide resource scarcity. This does not refer to multiple requesters competing for specific idle EEDs, because the edge orchestrator inherently prevents such device level contention through centralized management of EED assignments. The orchestrator maintains exclusive control over worker allocation, ensuring that each available EED is assigned to at most one requester at any time. Instead, congestion manifests mathematically through the effective idle worker intensity $\nu_{w_{idle}}$, which serves as a fundamental parameter in our spatiotemporal analysis and directly quantifies the system wide availability of computational resources. When $\nu_{w_{idle}}$ is low, indicating scarce EED availability, the allocation process becomes more challenging: requesters are forced to connect to more distant EEDs. This degradation reduces the D2D success probability, increases retransmission rates, and prolongs worker allocation time, thereby increasing the task response delay. To mitigate these effects, we tune the bias factor α to control the EEC-MEC split under congestion.

V. NUMERICAL RESULTS AND SIMULATION

This section presents numerical and simulation results to validate the proposed spatiotemporal models and evaluate their

TABLE III: Numerical Parameters

Parameter	Value
Workers Intensity (ν_w)	$7 \times 10^{-4} \text{ m}^{-2}$
Requester Intensity (ν_r)	$1 \times 10^{-4} \text{ m}^{-2}$
Carrier frequency	28 GHz
LoS and NLoS path loss exponent (α_L, α_N)	2, 4 [48]
Fading values for LoS and NLoS (N_L, N_N)	3, 2 [48]
Path loss intercepts (C_L, C_N)	-61.4 dB, -72 dB [48]
Main lobe gains ($M_w = M_r$)	5 dBi [49]
Side lobe gains ($m_w = m_r$)	-5 dBi [49]
SINR threshold (ξ)	5 dB [48]
3-dB beamwidth ($\theta_r = \theta_w$)	45° [44]
Bandwidth (B)	200 MHz [44]
Transmit power (P_t)	30 dBm [44]
Noise figure	10 dB [44]
Thermal noise power ($\sigma_{\text{th}}^2 = -174 \text{ dBm/Hz} + 10 \log_{10}(B[\text{Hz}]) + 10 \text{ dB (noise figure)}$) [49]	-81 dBm
Normalized noise $\sigma^2 = \sigma_{\text{th}}^2/P_t$	-111 dB
Task execution rate (μ_f)	0.02 task/second
D2D communication time (τ_c)	1 second
Maximum radius for LoS devices (R_L)	100 m
Reliability parameter (l)	3

performance across a wide range of operating conditions. We first confirm the accuracy of the analytical framework and then conduct sensitivity analysis that systematically explores the impact of key system parameters to assess the robustness of the observed trends. Unless otherwise stated, the default network parameters are listed in Table III. The key mmWave communication parameter values are selected in line with well-established mmWave studies in the literature. We emphasize that the developed mathematical model remains valid across a broad range of parameter values; the specific choices are adopted to demonstrate a typical operating conditions.

The Monte Carlo simulations are conducted over an area of 10 km^2 , where requesters and EED workers are generated as independent PPPs, and a typical requester is fixed at the origin. Around this requester, a disk of radius R_L defines the region of potential LoS links; devices inside this disk are tagged as LoS and those outside as NLoS. For every D2D link, we apply the distance-dependent path-loss model and sample channel fading gains from the corresponding Gamma distribution. Directional beamforming gain is computed from the actual link angles at the transmitter and receiver by selecting one of the four possible combined gains specified in Table II. The typical requester offloads to LoS workers within radius R_L according to the considered policy (random or location-aware); for each candidate worker, we compute its received SINR, including aggregate interference from LoS and NLoS requesters, and declare the offloading attempt successful if the SINR exceeds the threshold ξ . Repeating this over 10^5 independent network realizations yields the successful offloading probabilities of the first, second, and subsequent ordered workers, which are then converted into offloading rates.

Given a segmentation level n , we next simulate the ACTMC-based queueing model to obtain the end-to-end task response delay. The task is split into n equal segments, and we simultaneously track (i) sequential offloading, where an additional segment is assigned to a worker with an offloading rate determined following the procedure described previously,

and (ii) parallel execution of the offloaded segments, which complete with exponential service times of mean $1/(n\mu_f)$. Starting from an initial configuration where no segment has yet been offloaded, we track the time until all n segments have completed, which gives one realization of the end-to-end task response delay. Averaging these delays over 10^5 realizations yields the simulated average task response delay.

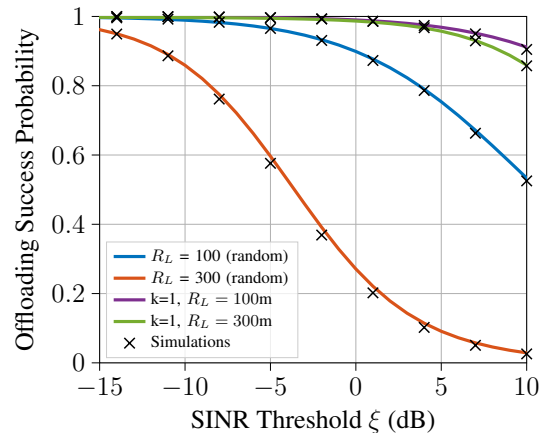


Fig. 3: Average D2D offloading success probability vs SINR threshold ξ .

Fig. 3 shows the average successful offloading probability p_s for the random and ordered EED-selection scenarios as a function of the SINR threshold ξ for different values of the LoS radius R_L . The case $k = 1$ corresponds to offloading to the nearest EED relative to the requester. The close match between the Monte Carlo simulations and the proposed analytical expressions confirms the accuracy of Theorems 1 and 3 over a wide range of ξ and R_L values. As expected, p_s decreases monotonically with ξ because a higher SINR threshold imposes a stricter link-quality requirement, making successful offloading less likely. The figure also reveals a strong sensitivity to the EED-selection strategy: offloading to the nearest device ($k = 1$) consistently achieves a much higher success probability than random selection, owing to the shorter average distance between the requester and the worker and, hence, a stronger received signal power. Furthermore, varying R_L illustrates the impact of the propagation environment. Larger R_L values increase the likelihood of longer LoS D2D links and simultaneously expose the requester to stronger interference from other LoS transmitters, both of which reduce the successful offloading probability.

Fig. 4 depicts the average task response delay as a function of the number of allocated workers (segments) n . The close agreement between the Monte Carlo simulations and the proposed analytical expressions validates Theorem 2 over a wide range of n values. Consistent with Fig. 3, selecting EEDs based on their distance from the requester significantly reduces the average response delay compared to random selection. This improvement stems from reduced offloading time, which results from the higher probability of successful offloading when associating with closer devices.

The figure also illustrates the sensitivity of the average

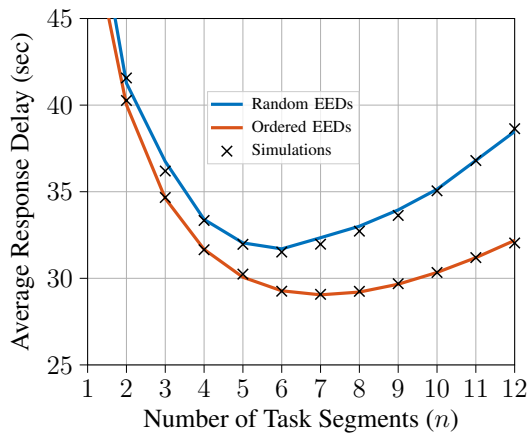


Fig. 4: Average task response delay vs the number of task segments.

delay to the segmentation level n , revealing a key design insight: an optimal number of segments n exists that minimizes the average task response delay. This behavior directly manifests the fundamental trade-off between communication and computation. For small to moderate n , increasing the number of segments reduces computation time because more workers share the task, while the additional communication overhead remains limited, thereby decreasing the overall delay. This trend continues until the communication delay, driven by longer links and more frequent retransmissions, becomes dominant and outweighs the reduction in computation delay, causing the average task response delay to increase.

A. Baseline Model Results

After validating the theoretical analysis against simulations, we conduct a sensitivity analysis to examine how key system parameters affect the performance of the baseline model.

Fig. 5 provides a sensitivity analysis of the optimal segmentation level with respect to key system parameters. In all subfigures, the red dots indicate the minimum average task response delay, i.e., the operating point where the system allocates the optimal number of workers n . In Fig. 5(a), we observe that the optimal number of workers strongly depends on the task execution rate μ_f . When μ_f is small, each worker is relatively slow, so involving more workers is beneficial to reduce the computation time, which shifts the optimal point to larger n . As μ_f increases, each worker can process tasks faster, and fewer workers are needed to minimize the response delay. Beyond the optimal n , adding more workers only increases the offloading overhead and the likelihood of retransmissions, which leads to performance degradation and explains the rise of the curves after the red dots.

Fig. 5(b) explores the impact of the ratio μ_f/λ_h , where λ_h is the average offloading rate. When μ_f is low relative to λ_h , the system is computation-limited: computation dominates the total delay, and more workers are required to reduce the response time. As μ_f increases relative to λ_h , the benefit of parallelism diminishes and the optimal n decreases. The case $\mu_f/\lambda_h = 1$ corresponds to a regime where communication and

computation delays are of the same order; in this case, offloading to a single worker is sufficient, as further segmentation would primarily increase communication costs without providing meaningful computation gains. Finally, Fig. 5(c) shows the effect of the SINR threshold ξ . Increasing ξ tightens the link-quality requirement, which reduces the successful offloading probability and increases the expected communication delay. In this communication-limited regime, it becomes preferable to use fewer workers to avoid excessive offloading overhead, and thus the optimal n shifts to smaller values. Overall, this figure highlights how the optimal segmentation level is highly sensitive to these main system parameters and provides practical guidelines for tuning n under different operating conditions.

To test the performance of our proposed EEC framework, we conduct a comparison against centralized MEC systems under various computational and congestion scenarios. The MEC employs an advanced parallel architecture [41], [50] where a single physical machine (PM) hosts multiple VMs for concurrent task processing. In all MEC scenarios, requesters within the LoS radius R_L offload complete tasks without partitioning, with congestion defined by the number of concurrent requesters served. We analyze a practical MEC that possesses computational power five times superior to that of an EED, which is consistent with typical values reported in literature [50]. We also analyze a more powerful MEC that possesses computational power ten times superior to that of an EED, representing a high-performance edge server, to further stress-test EEC's capabilities.

While practical MEC implementations suffer from I/O interference losses between VMs [41], our analysis conservatively assumes lossless parallelization where computational power is perfectly divided, presenting a best-case scenario for MEC performance. The MEC communication delay is computed similarly to that of the EEDs, while the average computation delay at the MEC is modeled as an exponential random variable with execution rates of $5\mu_f$ and $10\mu_f$ for the 5x and 10x power ratios, respectively, where $\mu_f = 0.007$. This establishes a true parallel-vs-parallel comparison: our distributed EEC architecture against a centralized but parallel MEC system.

Fig. 6 illustrates the average task response delay as a function of segmentation count (n) for EEC alongside MEC configurations. The results demonstrate that when ($\nu_{TMEC} = \nu_r$), EEC achieves significantly superior performance compared to the practical 5x MEC system, despite our ideal modeling of MEC capabilities. Remarkably, even when challenged by the more powerful 10x MEC configuration, EEC with optimal task segmentation ($n = 13$) maintains lower average task response delay. This optimal operating point represents a crucial balance between communication overhead and computational parallelism: below this threshold, EED computational resources remain underutilized, while beyond it, communication costs increase unnecessarily.

Furthermore, the analysis reveals MEC's inherent vulnerability to requester congestion and underscores EEC's core advantages stemming from distributed spatial parallelism. For the 5x MEC case, we observe three distinct performance regimes: minimal delay under no congestion ($\nu_{TMEC} = 0$), where the

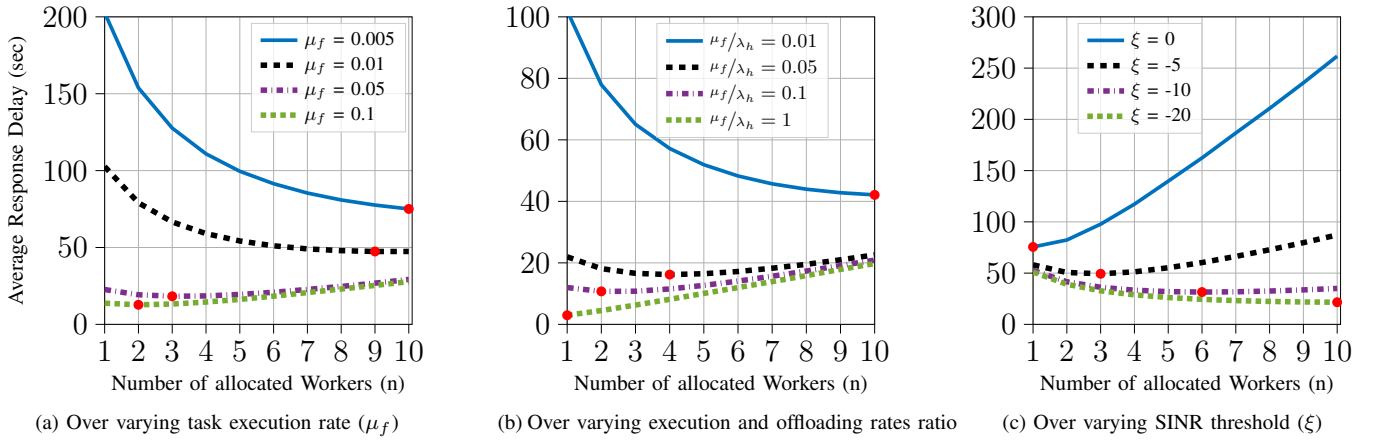


Fig. 5: Average task response delay (T_A) vs the number of allocated workers (n) for different system parameters.

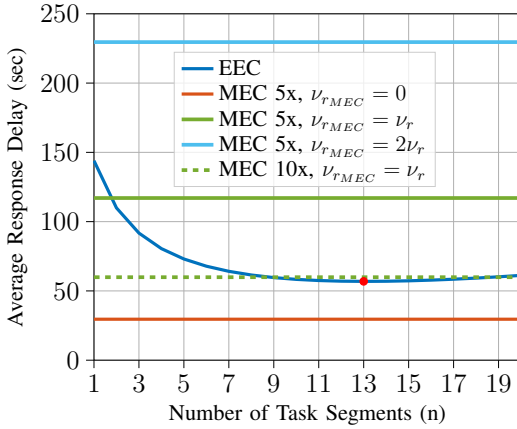


Fig. 6: MEC and EEC average response delay using varying MEC congestion scenarios.

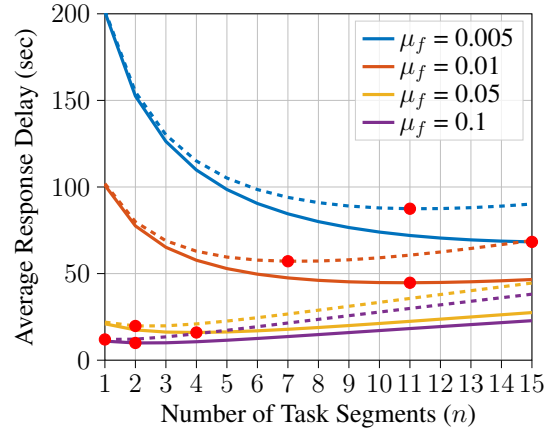


Fig. 7: Average response delay of random (dashed) versus ordered (solid) selection of devices under varying μ_f .

server dedicates full resources to a single task; significant degradation under moderate congestion ($\nu_{rMEC} = \nu_r$); and nearly doubled delay under high congestion ($\nu_{rMEC} = 2\nu_r$). This pronounced sensitivity to user load contrasts with EEC's congestion-resilient architecture, which effectively leverages underutilized edge resources to enable scalable, low-latency computation while avoiding the single-point congestion bottlenecks that afflict centralized MEC systems.

B. Advanced Model Results

Here, we conduct a sensitivity analysis of the advanced system model to investigate how key parameters affect performance and to derive system-level design insights.

Fig. 7 plots the average response delay across different task execution rates (μ_f), comparing random EED selection (dashed lines) with ordered selection (solid lines). Across all μ_f values, ordered selection achieves a lower average delay and shifts the delay-minimizing segmentation level to a larger optimal n (red markers). This gain, reaching up to about 22% delay reduction, is primarily driven by shorter offloading links, which increase the offloading success probability and reduce retransmissions, thereby lowering the effective communication cost and allowing the system to exploit parallelism more

efficiently. When combined with the low-overhead location acquisition methods discussed in Section IV-B, these results support the practicality of location-aware selection for improving EEC performance.

In Fig. 8, as expected, incorporating device failures increases the average task response delay. It also shifts the optimal number of task segments n to a larger value. This behavior stems from the coupling between the failure rate γ and the execution rate μ_f : when the task is divided into more segments, each worker handles a smaller portion of the task, which effectively reduces the per-segment failure rate γ_n and lowers the probability of failure events. However, this trend persists only up to a certain segmentation level. Beyond that point, further splitting the task requires allocating more devices, which significantly increases the offloading time; the resulting communication cost eventually dominates the reduction in failure probability and causes the average task response delay to rise. This figure therefore highlights the sensitivity of both the delay and the optimal segmentation level to the failure parameter γ and underscores the importance of jointly accounting for reliability and communication overhead when selecting n .

To examine congestion due to worker scarcity, we con-

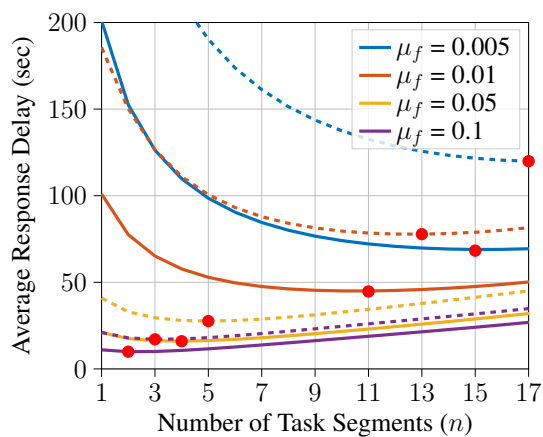


Fig. 8: Average task response delay, with EED failure (dashed) and without EED failure (solid), for varying task execution rates (μ_f).

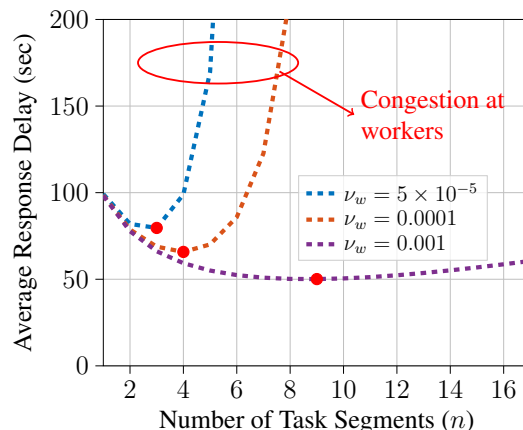


Fig. 9: Average response delay over different ν_w values under the ordered offloading while considering failure.

duct an additional experiment to evaluate the impact of the worker intensity ν_w on the average task response delay. As shown in Fig. 9, lower values of ν_w lead to significantly higher delays and a smaller optimal number of task segments. This behavior is primarily due to the limited number of available LoS EEDs, which induces congestion as multiple requesters compete for a scarce set of workers. In such cases, the few accessible workers are often located farther from the requester, resulting in weaker D2D links and a reduced offloading success probability, thereby increasing the time required to allocate each task segment. The figure further shows that, under low ν_w , the average response delay rises sharply beyond a certain segmentation level and continues to increase as n grows, reflecting the growing difficulty of recruiting reliable LoS workers. Consequently, in worker-congested regimes, deviations from the optimal segmentation level incur a much larger delay penalty. This sensitivity to ν_w highlights the importance of the EEC-MEC collaboration approach proposed in this work: by offloading excess demand to MEC resources when worker intensity is low, the system can effectively mitigate congestion at the EED level and maintain low response delays under constrained worker availability.

To guide system design, Fig. 10 plots the contour of the optimal task segmentation, n^* , as a joint function of worker intensity (ν_w) and task execution rate (μ_f). This design map reveals three core operational regimes. First, in the high ν_w and low μ_f regime, n^* is high. This is because abundant nearby workers minimize the communication overhead of offloading, while the slow task execution rate ensures that the computation speedup from parallelization is significant enough to outweigh these communication costs. Second, under low ν_w , n^* is forced to be low regardless of μ_f , confirming that worker scarcity creates a communication bottleneck that dominates performance. Third, for high μ_f , n^* remains low even with ample workers, as the minimal computation time makes parallelization gains insignificant compared to the fixed offloading cost. Collectively, this map provides a vital practical tool, showing that optimal operation requires matching the segmentation strategy to the specific environment. Aggres-

sive parallelization is only beneficial in dense networks with computationally intensive tasks, while simpler tasks or sparse networks require minimal segmentation to avoid excessive communication overhead.

Fig. 11 illustrates the task completion probability as a function of the EED reliability parameter l . As expected, the completion probability decreases as l decreases, reflecting a higher likelihood of worker failure and, consequently, a reduced chance of successful task execution. More importantly, increasing the number of task segments n consistently enhances the completion probability. This improvement arises because smaller segments have shorter execution durations, which reduces the risk that a worker fails before finishing its assigned portion. However, beyond a certain segmentation threshold, the marginal gains in reliability begin to diminish. As segment sizes become very small, further segmentation provides limited additional benefit while incurring extra communication overhead, which can increase the total task response delay. Hence, achieving very high reliability (e.g., a completion probability of 0.99) may require operating at a segmentation level n that is larger than the value minimizing the average response delay. Such scenarios often arise in applications with stringent reliability requirements, where the system designer may deliberately choose a higher n to ensure task completion, even at the expense of increased latency.

These results highlight the value of considering task completion probability as a reliability metric alongside the average task response delay. Together, these metrics provide a more complete perspective for selecting a segmentation level n that balances latency and reliability for a given application. It is also worth noting that the system congestion level plays a critical role in this trade-off. As discussed earlier, while increasing n generally improves reliability, the associated delay penalty is highly sensitive to congestion. For example, in highly congested scenarios with low worker intensity, pursuing very high reliability by increasing n can cause a sharp rise in response delay, as clearly illustrated in Fig. 9.

Next, we investigate EEC-MEC collaboration and how the optimal α systematically shifts in response to changing con-

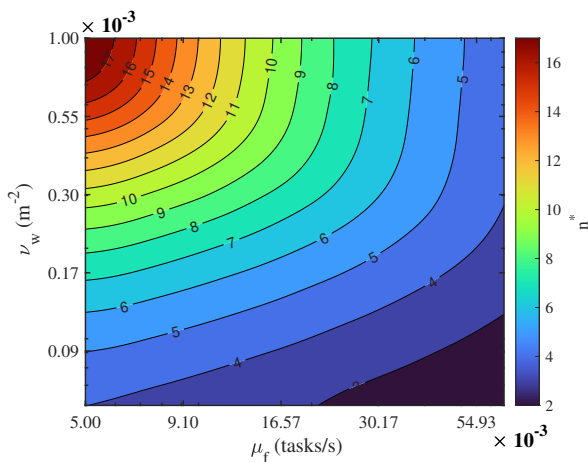


Fig. 10: The optimal number of task segments (n^*) as a joint function of worker intensity (ν_w) and task execution rate (μ_f).

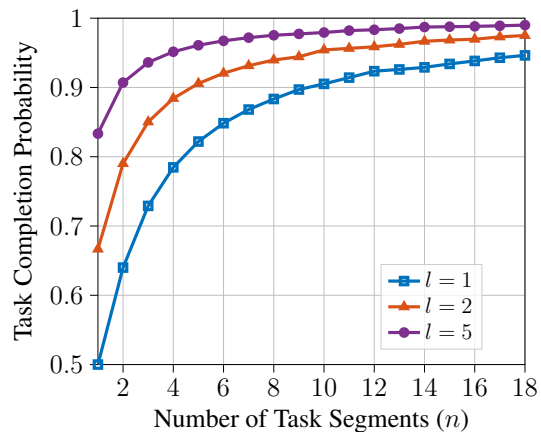
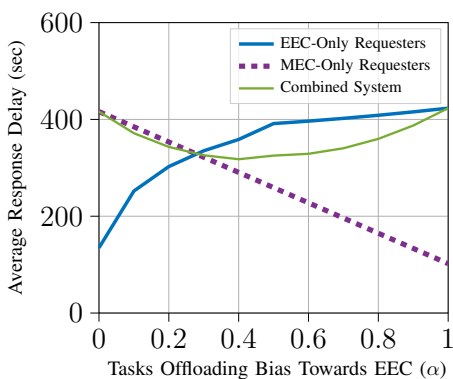
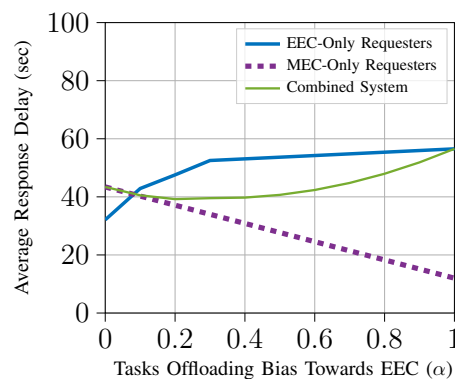


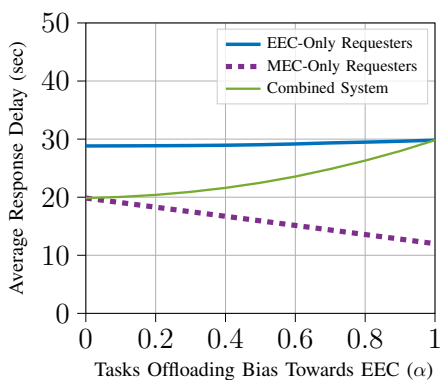
Fig. 11: Task completion probability over varying EEDs reliability Parameters



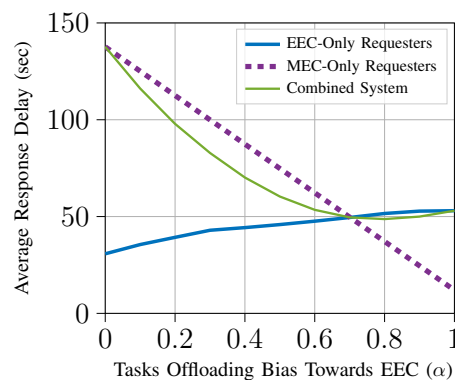
(a) Low task execution rate $\mu_f = 0.002$



(b) Low worker intensity $\nu_w/4$, $\mu_f = 0.02$



(c) Low requester intensity $\nu_r/4$, $\mu_f = 0.02$



(d) High requester intensity $4\nu_r$, $\mu_f = 0.02$

Fig. 12: Average task response delay vs. portion of requesters offloading to EED (α) under different system parameters.

gestion conditions, quantified through the effective idle worker intensity $\nu_{w_{idle}}$. Using the ordered offloading with MEC computational power set to five times that of a single EED, Fig. 12 depicts the average response delay for: the portion of requesters using EEC exclusively, the portion utilizing MEC exclusively, and the combined system across varying bias factor α values. The combined system represents the overall

performance metric we aim to optimize, reflecting the average response delay experienced by any typical requester in the system. The depicted scenario begins with a single requester utilizing EEC, while the others offload to MEC. Subsequently, as α increases, the situation gradually shifts until one requester exclusively relies on MEC and the rest use EEC.

Fig. 12(a) shows the average response delay under low task

execution rate, which inherently reduce idle worker density and contribute to system congestion. The low execution rate prolongs task completion times, decreasing the effective idle worker intensity $\nu_{w_{idle}}$ and constraining the available EED pool. This congestion effect reveals an optimal operating point at $\alpha = 0.4$, where the system balances EEC utilization against resource scarcity. As α increases from zero, growing EEC usage alleviates MEC congestion, initially reducing combined system delay. However, beyond $\alpha = 0.4$, the limited LoS EED availability becomes critically strained, forcing allocations to more distant workers and degrading D2D performance, ultimately increasing the average response delay. The optimal α thus represents the precise balance that minimizes system-wide delay by avoiding both excessive MEC load (at low α) and EEC resource saturation (at high α).

Fig. 12(b) examines the impact of low worker intensity, revealing how fundamental resource scarcity shifts the optimal operating point to $\alpha = 0.2$. The critically low worker density dramatically reduces $\nu_{w_{idle}}$ and limits the availability of LoS EEDs. This scarcity forces the allocation of more distant workers, increasing communication delays and retransmissions. The optimal α shift to 0.2 represents the system's adaptation to this constrained environment, where only minimal EEC utilization can be supported without overwhelming the limited worker pool and degrading overall performance.

Fig. 12(c) presents the average response delay under low requester intensity, which significantly increases $\nu_{w_{idle}}$ and reduces system-wide congestion. The optimal operating point at $\alpha = 0$ demonstrates that exclusive MEC offloading minimizes average response delay. In this uncongested regime, MEC's superior computational power provides faster task completion than distributed EEC processing, despite the availability of LoS EEDs. While EEC performance remains stable across α values due to ample worker availability, the absence of requester-level congestion eliminates MEC's scalability limitations, allowing its computational advantage to dominate. This result validates our framework's ability to identify when centralized MEC resources outperform distributed EEC capabilities based on system congestion conditions quantified through $\nu_{w_{idle}}$.

Fig. 12(d) presents a high requester intensity scenario, where the optimal operating point shifts dramatically to $\alpha = 0.8$. This significant result demonstrates that extensive EEC utilization becomes the dominant strategy under high congestion, as its distributed nature provides crucial scalability that centralized MEC cannot match when overwhelmed by high requester density. The optimal $\alpha = 0.8$ indicates that 80% of requesters should be served by EEC resources, leveraging the parallel processing capabilities of EEDs to alleviate MEC congestion. This finding clearly validates our paper's core contribution: EEC emerges as a vital computational paradigm in high-density scenarios, where its distributed architecture and spatial resource pooling overcome the scalability limitations of traditional edge computing.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel spatiotemporal framework for EEC in large-scale mmWave networks, integrating SG with

an ACTMC to jointly model mmWave D2D offloading and parallel computation, including their temporal overlap. The framework enables a tractable end-to-end evaluation of two key metrics, the average task response delay and the task completion probability, providing a unified view of latency and reliability. Our analysis, validated by Monte Carlo simulations and sensitivity studies, reveals a fundamental communication-computation trade-off. This trade-off yields an optimal task segmentation level that minimizes delay, balancing the benefits of parallelism against the overhead of excessive offloading. This optimum depends critically on operating conditions, such as D2D link quality, computation speed, and worker density, making aggressive parallelization beneficial for computation-heavy tasks with sufficient nearby workers. Furthermore, location-aware EED selection consistently outperforms random selection; by improving offloading success, it pushes the delay-optimal segmentation level higher, enabling greater parallelism gains.

Extending the analysis to practical impairments, we find that EED failures shift the delay-optimal segmentation level higher, as smaller segments enhance resilience against device failure. Conversely, worker scarcity reduces the optimal segmentation level to limit offloading overhead; in such resource-constrained scenarios, deviating from the optimum incurs a severe delay penalty. Furthermore, meeting stringent reliability targets may require operating above the delay-optimal point, underscoring the trade-off between latency and reliability. Finally, regarding EEC-MEC collaboration, we demonstrate that the optimal bias factor adapts to congestion by balancing MEC load against worker scarcity to minimize system delay. Overall, this work provides a rigorous analytical foundation and practical, sensitivity-aware guidelines for EEC system design, enabling informed decisions on task segmentation, offloading strategies, and collaboration across a broad range of operating conditions.

Future work will generalize the proposed spatiotemporal framework to support heterogeneous segment sizes and to incorporate dependent-task workflows with precedence constraints, such as DAG-structured workflows, including synchronization requirements. We plan to explore two complementary directions: (i) a stage-wise abstraction that represents the workflow as an ordered sequence of dependency-constrained stages, where the proposed analysis is applied at the stage level and the end-to-end delay is obtained by composing stage delays; and (ii) a partition-based abstraction that groups dependent subtasks into a small number of macro-tasks executed under precedence constraints, together with corresponding readiness-aware extensions to the delay characterization. In parallel, heterogeneous EED capabilities will be addressed by incorporating selection policies that map segments or workflow components to workers according to their computational power. Finally, we also plan to extend the framework to applications with non-negligible result payloads. This requires augmenting the ACTMC with an explicit result-return phase, and studying how result communication influences segmentation decisions, location-aware offloading benefits, and the MEC-EEC load-balancing bias factor.

REFERENCES

[1] M. Abdelhadi, S. Sorour, H. ElSawy, S. A. Elsayed, and H. Hassanein, "Parallel computing at the extreme edge: Spatiotemporal analysis," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 5692–5698.

[2] X. Ma, X. Liu, N. Ansari, and J. Chang, "Artificial Intelligence of Things in 6G Networks: Unlocking Opportunities and Overcoming Challenges," *IEEE Commun. Mag.*, 2025.

[3] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 905–974, 2023.

[4] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, 2021.

[5] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2021.

[6] A. Hazra, P. Rana, M. Adhikari, and T. Amgoth, "Fog computing for next-generation internet of things: fundamental, state-of-the-art and research challenges," *Comput. Sci. Rev.*, vol. 48, p. 100549, 2023.

[7] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge computing with artificial intelligence: A machine learning perspective," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.

[8] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[9] S. Bagchi, M.-B. Siddiqui, P. Wood, and H. Zhang, "Dependability in edge computing," *Commun. ACM*, vol. 63, no. 1, pp. 58–66, 2019.

[10] J. Portilla, G. Mujica, J.-S. Lee, and T. Riesgo, "The extreme edge at the bottom of the internet of things: A review," *IEEE Sensors J.*, vol. 19, no. 9, pp. 3179–3190, 2019.

[11] R. F. El-Khatib, S. A. Elsayed, N. Zorba, and H. S. Hassanein, "Proactive Task Allocation in Extreme Edge Computing for Digital Twin Services," *IEEE Internet Things J.*, 2025.

[12] H.-S. Lee and J.-W. Lee, "Task offloading in heterogeneous mobile cloud computing: Modeling, analysis, and cloudlet deployment," *IEEE Access*, vol. 6, pp. 14908–14925, 2018.

[13] I. M. Amer, S. M. Oteafy, S. A. Elsayed, and H. S. Hassanein, "Task provisioning in unreliable edge networks: Inferring utility," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 3015–3020.

[14] H. Wang, Q. Li, H. Kang, D. Hu, L. Ma, G. Tyson, Z. Yuan, and Y. Jiang, "Paraloupe: Real-time video analytics on edge cluster via mini model parallelization," *IEEE Trans. Mobile Comput.*, 2024.

[15] Z. Zhao, K. M. Barijough, and A. Gerstlauer, "Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, no. 11, pp. 2348–2359, 2018.

[16] M. M. Razaq, B. Tak, L. Peng, and M. Guizani, "Privacy-aware collaborative task offloading in fog computing," *IEEE Trans. Comput. Social Syst.*, vol. 9, no. 1, pp. 88–96, 2021.

[17] Y. Yang, Y. Shi, C. Yi, J. Cai, J. Kang, D. Niyato, and X. Shen, "Dynamic Human Digital Twin Deployment at the Edge for Task Execution: A Two-Timescale Accuracy-Aware Online Optimization," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12262–12279, 2024.

[18] J. Chen, C. Yi, S. Gong, H. Du, W. Wu, J. Kang, and D. Niyato, "Generative AI-Aided QoE-Aware Resource Allocations for RIS-Assisted Digital Twin Interaction with Uncertain Evolution," *IEEE Trans. Mobile Comput.*, pp. 1–18, 2025.

[19] A. Mohammadzadeh, M. Masdari, F. S. Gharehchopogh, and A. Jafarian, "Improved chaotic binary grey wolf optimization algorithm for workflow scheduling in green cloud computing," *Evol. Intell.*, vol. 14, no. 4, pp. 1997–2025, 2021.

[20] A. Mohammadzadeh, M. Masdari, and F. S. Gharehchopogh, "Energy and cost-aware workflow scheduling in cloud computing data centers using a multi-objective optimization algorithm," *J. Netw. Syst. Manag.*, vol. 29, no. 3, p. 31, 2021.

[21] L. Li, C. Zhou, P. Cong, Y. Shen, J. Zhou, and T. Wei, "Makespan and security-aware workflow scheduling for cloud service cost minimization," *IEEE Trans. Cloud Comput.*, vol. 12, no. 2, pp. 609–624, 2024.

[22] H. Jiang, X. Dai, Z. Xiao, and A. K. Iyengar, "Joint task offloading and resource allocation for energy-constrained mobile edge computing," *IEEE Trans. Mobile Comput.*, 2022.

[23] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 1451–1455.

[24] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, 2015.

[25] C. Yi, J. Cai, and Z. Su, "A Multi-User Mobile Computation Offloading and Transmission Scheduling Mechanism for Delay-Sensitive Applications," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 29–43, 2020.

[26] H. Cao and J. Cai, "Distributed Multiuser Computation Offloading for Cloudlet-Based Mobile Cloud Computing: A Game-Theoretic Machine Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 752–764, 2018.

[27] K. Moghaddasi, S. Rajabi, and F. S. Gharehchopogh, "Multi-objective secure task offloading strategy for blockchain-enabled IoV-MEC systems: a double deep Q-network approach," *IEEE Access*, vol. 12, pp. 3437–3463, 2024.

[28] A. M. Rahmani, A. Haider, K. Moghaddasi, F. S. Gharehchopogh, K. Aurangzeb, Z. Liu, and M. Hosseinzadeh, "Self-learning adaptive power management scheme for energy-efficient IoT-MEC systems using soft actor-critic algorithm," *Internet Things*, vol. 31, p. 101587, 2025.

[29] S. B. Azmy, N. Zorba, and H. S. Hassanein, "Incentive-vacation queuing in extreme edge computing: An analytical reward-based framework," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2183–2195, 2024.

[30] M. Masoumi, E. Carmona-Cejudo, I. de Miguel, C. Torres-Pérez, and R. J. D. Barroso, "Dynamic Joint Scheduling of Movement and Data Processing Tasks using Extreme-Edge Computing in Multi-AGV Scenarios," *IEEE Open J. Ind. Electron. Soc.*, 2025.

[31] Z. Safavifar, E. Gyamfi, E. Mangina, and F. Golpayegani, "Multi-objective deep reinforcement learning for efficient workload orchestration in extreme edge computing," *IEEE Access*, vol. 12, pp. 74558–74571, 2024.

[32] Z. Safavifar, C. Machalikh, J. Xie, and F. Golpayegani, "Sustainable dependent sub-tasks orchestration at extreme edge computing: A partitioning-based deep reinforcement learning approach," *ACM J. Comput. Sustainable Soc.*, vol. 3, no. 2, pp. 1–31, 2025.

[33] G. Drainakis, P. Pantazopoulos, K. V. Katsaros, V. Sourlas, T. Xirofotos, N. Baganal-Krishna, A. Rizk, R. Horvath, G. Scivoletto, A. Amditis, *et al.*, "Service Orchestration at the Extreme-Edge: An Experimental Investigation Over a 5G Testbed," in *ICC 2025-IEEE International Conference on Communications*. IEEE, 2025, pp. 844–849.

[34] M. Elbayoumi, W. Hamouda, and A. Youssef, "Edge computing and multiple-association in ultra-dense networks: Performance analysis," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5098–5112, 2022.

[35] Q. Cheng, G. Cai, J. He, and G. Kaddoum, "Design and performance analysis of MEC-aided LoRa networks with power control," *IEEE Trans. Veh. Technol.*, 2024.

[36] M. Song, H. H. Yang, H. Shan, J. Lee, and T. Q. Quek, "Age of information in wireless networks: Spatiotemporal analysis and locally adaptive power control," *IEEE Trans. Mobile Comput.*, vol. 22, no. 6, pp. 3123–3136, 2021.

[37] Y. Nabil, H. ElSawy, S. Al-Dharrab, H. Mostafa, and H. Attia, "Data Aggregation in Regular Large-Scale IoT Networks: Granularity, Reliability, and Delay Tradeoffs," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17767–17784, 2022.

[38] Y. Chen, J. Liu, and P. Siano, "SGedge: Stochastic geometry-based model for multi-access edge computing in wireless sensor networks," *IEEE Access*, vol. 9, pp. 111238–111248, 2021.

[39] S.-W. Ko, K. Han, and K. Huang, "Wireless networks for mobile edge computing: Spatial modeling and latency analysis," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5225–5240, 2018.

[40] Y. Gu, Y. Yao, C. Li, B. Xia, D. Xu, and C. Zhang, "Modeling and analysis of stochastic mobile-edge computing wireless networks," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 14051–14065, 2021.

[41] M. Emara, H. ElSawy, M. C. Filippou, and G. Bauch, "Spatiotemporal dependable task execution services in MEC-enabled wireless systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 211–215, 2021.

[42] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1038–1051, 2020.

[43] Y. Gu, C. Yin, Y. Guo, B. Xia, and Z. Chen, "Communication-computation-aware user association in MEC HetNets: A meta-analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8919–8933, 2023.

[44] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhatieb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, 2016.

- [45] Y. Yang, M. Chen, Y. Blankenship, J. Lee, Z. Ghassemlooy, J. Cheng, and S. Mao, "Positioning using wireless networks: Applications, recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, 2024.
- [46] A. Musa, J. Biagioni, and J. Eriksson, "Trading off accuracy, timeliness, and uplink usage in online gps tracking," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 2124–2136, 2015.
- [47] Y. Nabil, H. ElSawy, and H. S. Hassanein, "System-Level Analysis of Dual-Mode Networked Sensing: ISAC Integration and Coordination Gains," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 7970–7987, 2026.
- [48] X. Yu, J. Zhang, M. Haenggi, and K. B. Letaief, "Coverage analysis for millimeter wave networks: The impact of directional antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1498–1512, 2017.
- [49] J. Chen, X. Ge, and Q. Ni, "Coverage and handoff analysis of 5G fractal small cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1263–1276, 2019.
- [50] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato, and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1359–1374, 2019.



Yasser Nabil (Graduate Student Member, IEEE) received the B.Sc. degree in electronics and communications engineering from Alexandria University, Alexandria, Egypt, and the M.Sc. degree in electronics and communications engineering from Cairo University, Giza, Egypt. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Queen's University, Kingston, ON, Canada. His research interests include the integration of sensing and communication, wireless communication systems and their statistical modeling, the Internet of

Things, non-terrestrial networks, and edge computing. His work has appeared in leading IEEE journals and conferences, including the IEEE Transactions on Wireless Communications, the IEEE Internet of Things Journal, the IEEE Transactions on Vehicular Technology, and IEEE ICC.



Mahmoud Abdelhadi holds a Bachelor of Science degree in Computer Science from the Applied Science University in Amman, Jordan, and a Master of Science in Computer Science from Queen's University in Kingston, Ontario. During his graduate studies, he served as a Research Assistant in the RTL Lab at Queen's University, where he contributed to research in networks and edge computing. His academic background encompasses software systems, algorithms, and 5G networks. Currently, he works as a Software Engineer at Amazon, where he focuses

on designing and implementing scalable software solutions. His professional interests include cloud computing, systems architecture, and the development of high-performance, fault-tolerant applications.



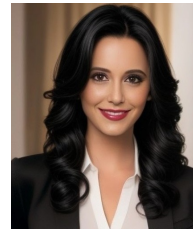
Sameh Sorour (Late). Dr. Sorour was an assistant professor at the School of Computing, Queen's University, where he led several projects on autonomous and connected vehicles, edge intelligence, and wireless networks and services. He was a senior IEEE member and an Editor for IEEE Communications Letters. His research and educational interests lie in the broad areas of advanced computing, learning, and networking technologies for cyber-physical and autonomous systems. Dr. Sorour died in 2021. He was a wonderful scholar, instructor, mentor, and

colleague; he will be greatly missed.



Hesham ElSawy (Senior Member, IEEE) an Associate Professor with the School of Computing, Queen's University, Kingston, ON, Canada. Prior to that, he was an assistant professor at King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, a Post-Doctoral Fellow at the King Abdullah University of Science and Technology (KAUST), Saudi Arabia, a Research Assistant at TRTech, Winnipeg, MB, Canada. He received the Ph.D. degree in electrical engineering from the University of Manitoba, Canada, in 2014. He conducts

research in the broad area of wireless communications and networking with a special focus on 5G/6G networks, Internet of Things, edge computing, non-terrestrial networks, and wireless security. Dr. ElSawy is a recipient of the IEEE ComSoc Outstanding Young Researcher Award for Europe, Middle East, and Africa Region in 2018. He also received several best paper awards including the IEEE COMSOC Best Tutorial Paper Award in 2020 and IEEE COMSOC Best Survey Paper Award 2017. He is an Editor of the IEEE Transactions on Wireless Communications, the IEEE Transactions on Network Science and Engineering, and the IEEE Communications Letters.



Sara A. Elsayed received her PhD in Computer Science from the School of Computing at Queen's University, Canada, in 2020. She is currently an Assistant Professor in the Department of Computer Science at the University of Calgary. Previously, Elsayed worked as a Postdoctoral Fellow and Adjunct Assistant Professor at Queen's University, where she also managed and coordinated a large-scale research project focused on democratizing edge computing and edge intelligence in collaboration with Distributive, Ltd. In 2023, she received the Queen's School

of Computing (QSC) research award in recognition of her research contributions. Prior to joining Queen's University, Elsayed was an Assistant Lecturer in the Faculty of Computers and Information, Cairo University, Egypt. She was the Co-founder and Program Manager of the Teaching & Instructional Center (TIC), and was an Assistant Focal Point in the Support Office for Research Cooperation & Mobility (SORCAM), Engineering Sector, Egypt. She is also a certified Cisco Networking Academy Instructor. Her research interests include Edge Computing, Edge Intelligence, Vehicular Networks, Caching, Internet of Things, Artificial Intelligence, and Intelligent Systems. She has several publications in top venues and is a member of the IEEE.



Hossam S. Hassanein (Fellow, IEEE) is currently a leading Researcher in the areas of broadband, wireless and mobile networks architecture, protocols, control, and performance evaluation. His record spans more than 600 publications in journals, conferences, and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. He has received several recognition and best paper awards at top international conferences. He is the Founder and the Director of the Telecommunications Research Laboratory (TRL), School of Computing, Queen's

University, with extensive international academic and industrial collaborations. He is a recipient of the 2016 IEEE Communications Society Communications Software Technical Achievement Award for outstanding contributions to routing and deployment planning algorithms in wireless sensor networks and the 2020 IEEE IoT, Ad Hoc and Sensor Networks Technical Achievement and Recognition Award for significant contributions to technological advancement of the Internet of Things, ad hoc networks, and sensing systems. He is the former Chair of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). He is an IEEE Communications Society Distinguished Speaker [a Distinguished Lecturer (2008–2010)].