GRANT: Genetic-based RAN orchestration Tuning for latency sensitive XR verticals

Luca Mastrandrea[1], Alessandro Priviero[1],Ioannis Chatzigiannakis[2], Stefania Colonnese[1]
[1]*Department of Information Engineering, Electronics and Telecommunications*
[2]*Department of Computer, Control, and Management Engineering*
Sapienza University of Rome, Italy

***Abstract*—In next-generation networks, verticals such as industrial automation, remote surgery, and first-responder training require ultra-low latency mobile services, often coupled with high throughput demands, particularly in the context of extended reality applications. To meet these requirements, we propose a genetic algorithm-based orchestration framework for Radio Access Network (RAN) resource management, which leverages collaborative scheduling across multiple base stations and is specifically designed for extended reality scenarios.**

We define a latency-sensitive cost function for resource allocation based on a graph-based service model, and introduce the Genetic-based RAN Orchestration Tuning (GRANT) framework to optimize resource distribution. This approach effectively controls orchestration complexity while requiring only minimal knowledge of user subscription data.

Numerical simulations demonstrate that the proposed method outperforms existing state-of-the-art solutions, offering enhanced quality of experience, particularly for latency-critical services such as extended reality.

***Index Terms*—5G NR, Edge Computing, Extended Reality, Joint Allocation, Cooperative Scheduling.**

I. Introduction

5G technologies enable high-throughput latency-sensitive mobile services, such as extended reality (XR), into diverse fields, supporting interactive immersive multimedia experiences [1] beyond entertainment to educational uses [2], [3]. The effectiveness of immersive multimedia depends on high data throughput and minimal latency to maintain a satisfactory Quality of Experience [4], and XR applications often require low-latency synchronized interactions among multiple users, potentially leading to network congestion at the radio access [5]. Hence, recent studies have explored radio resources allocation strategies for users covered by multiple radio access points. Overlapping radio coverage among access points provides advantages for resource allocation [6]–[8] thanks to cooperative radio resource scheduling [9], [10] techniques.

We address the problem of resource allocation for users under the coverage of different Radio Access Network (RAN) base stations, and we introduce GRANT (Genetic-based RAN Orchestration Tuning), an optimization approach tailored for XR latency-sensitive verticals. Genetic algorithms have found application to Mobile Edge Computing (MEC) resources allocation [11], content distribution in fog computing systems [12]. Herein, we resort to genetic algorithms to solve the bandwidth allocation problem for minimizing service delay in cooperative RAN resource scheduling. The literature has addressed in the past the case of non-cooperative scheduling, and the minimum delay disjoint resource allocation has been solved analytically in [13]. Recently, the case of guaranteed fluidity and the case of minimum delay using joint cooperative scheduling has been tackled by greedy algorithms [10], [14].

The GRANT algorithm builds upon cooperative capabilities offered by 5G network function virtualization at the RAN and leads to the joint minimum delay allocation by genetic stochastic optimization [9], [15]–[17]. Specifically, GRANT provides an efficient method for solving optimization problems while maintaining control over the orchestration computational complexity based on algorithm parameters. Moreover, GRANT is highly parallelizable, ensuring scalability with the number of RAN users. Faster convergence compared to other optimization methods, coupled with controlled memory utilization, also leads to reduced energy consumption [18].

The main contributions of this work are as follows: formulating the minimum delay RAN resource allocation problem under cooperative scheduling, we apply GRANT to optimize the non-differentiable function governing the cooperative scheme. To reduce the search space, we initialize GRANT with the solution to the minimum delay problem in the non-cooperative case. Numerical simulations show that

This work was partially supported by the European Union's Horizon HADEA research and innovation program under grant Agreement 101092851 XR2LEARN project and by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP B53C22004050001, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

ISBN 978-3-903176-72-0 ©2025 IFIP

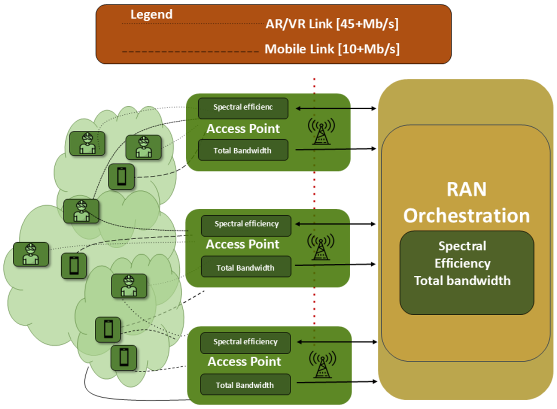


Fig. 1: Mobile latency-sensitive vertical scenario: a RAN Intelligent Controller orchestrates the radio access network (RAN) resources of multiple access points with overlapping coverage, allowing users to benefit from cooperative scheduling.

our approach outperforms state-of-the-art solutions in 3GPP-compliant realistic scenarios by leveraging the flexibility of next-generation networks.

The paper is organized as follows: Section II reviews the system architecture. Section III outlines the orchestration policy and presents the GRANT method. Section IV provides numerical simulations demonstrating the superiority of GRANT over current solutions, and discusses algorithmic and implementation challenges, while Sec. V concludes the paper and outlines future research directions.

II. SYSTEM ARCHITECTURE

The scenario considers a set of N users served by M partially overlapping base stations, whose resources are jointly managed by a RAN Intelligent Controller, as described in [14].

We assume that the execution of an XR application is divided into time slots of equal duration. In the τ-th time slot, the n-th user transmits or receives packets belonging to different traffic types, such as AR, VR, CG, and mobile video. Let $\theta_n[\tau]$ denote the throughput requested by user n during time slot τ, which includes both multimedia data and auxiliary control signals, such as audio and gesture data. Without loss of generality, we focus on resource allocation within a single time slot. Therefore, we omit the time index τ for simplicity in the remainder of the discussion.

Each user requires a minimum throughput $\theta_n^{(min)}$, determined by the specific XR application in use, and imposes a maximum allowable delay $\delta_n^{(max)}$ on the transmission of packets, to meet application-layer quality of service requirements. In general, delay comprises both RAN and possible mobile edge computing components; however, in this study we focus specifically on the RAN component of the delay experienced by user n, which is related to the throughput demand θ_n.

Let M denote the number of base stations coordinated by the joint resource orchestrator, and let $B^{(m)}n$ represent the available bandwidth at base station m allocated to user n. We define a spectral efficiency matrix ARAN, whose (m, n)-th element, $\eta_n^{(m)}$, denotes the spectral efficiency of the wireless channel between base station m and user n.

The delay experienced at the RAN by the n-th user for the transmission of a packet of size θ_n bits is computed as

$$\delta_n^{(RAN)} = \frac{\theta_n}{\sum_{m=0}^{M-1} B_n^{(m)} \cdot \eta_n^{(m)}} \tag{1}$$

whereas the overall delay, including the delay introduced by the MEC facilities is higher and computed as $\delta_n = \delta_n^{(RAN)} + \delta_n^{(MEC)}$.

We aim to optimize the RAN resource allocation:

$$\mathbf{b} = \left[B_0^{(0)}, \cdots, B_{N-1}^{(0)}, \cdots, B_0^{(M-1)}, \cdots, B_{N-1}^{(M-1)}\right]^T,$$

so as to minimize the overall RAN delay

$$\min_{\{B_n^{(m)}\}} \sum_{n=0}^{N-1} \delta_n^{(RAN)} \tag{2}$$

under the constraints over the actual base stations' bandwidths: $\sum_{n=0}^{N-1} B_n^{(m)} \leq B_{TOT}^{(m)},\ m=0, \cdots M-1$.

The problem of minimum RAN scheduling delay within a cooperative framework can be expressed in terms of a linear problem [10] as follows:

$$\begin{cases} \mathcal{A}\mathbf{b} = \mathcal{B}, \text{ subject to:} \\ \sum_{n=0}^{N-1} B_n^{(m)} \leq B_{TOT}^{(m)},\ m \in \{0, \dots, M-1\} \end{cases}$$

where the matrix $\mathcal{A}$ depends on the spectral efficiency matrix A_{RAN} and the known term $\mathcal{B}$ is set based on the maximum tolerated users' delays $\delta_0^{(\text{MAX})}, \cdots \delta_{N-1}^{(\text{MAX})}$, using: $\mathcal{B}_n \geq \frac{\theta_n}{\delta_n^{(\text{MAX})}}$, $n = 0, \cdots N-1$. Herein, we address the optimization in (2) by resorting to a genetic algorithm approach, as described in the following.

III. THE GENETIC RAN TUNING (GRANT)

We propose a stochastic optimization genetic algorithm inspired to biological evolutionary processes, referred to as Genetic RAN Tuning (GRANT), to minimize overall delay by solving the cooperative bandwidth allocation problem on a slotted-time basis. GRANT-based cooperative scheduling is a smart, flexible evolution of the disjoint scheduling enabled by 5G technologies. The presence of multiple radio access resources increases the complexity with respect to the disjoint resource allocation scenario addressed in [13], and has been previously approached using game-theoretic methods [14] and greedy algorithms [10]. The output of the proposed optimization can serve as input to lower-layer cooperative scheduling strategies, such as those in [15], which operate at finer MAC-layer time scales.

GRANT belongs to the class of genetic algorithms, i.e. stochastic optimization methods that minimize an objective (fitness) function f over a set of variables referred to as genes. In our problem formulation, each gene represents the amount of resources allocated by a station to a specific user, reflecting the area coverage constraints: if a station cannot serve a user, the corresponding gene is constrained to zero, thereby preventing any allocation and reducing the number of genes to be optimized. An individual is represented by an array of $N \times M$ genes. The initial population of N_i individuals is generated using a low complexity, minimum delay, non cooperative allocation, as proposed in [13]. Subsequently, N_G new generations are formed by genetic recombination of the fittest individuals and random gene modifications, governed by crossover and mutation probabilities p_c and p_m, respectively. The mutation coefficient μ controls the extend of the random modifications. At each generation, the individuals with the highest fitness function values are retained. This process ensures broad exploration of the search space, although it may occasionally result in suboptimal solutions.

The fitness function is defined as the cumulative RAN delay, as expressed in (2). To prevent constraint violations, a penalty factor is introduced to eliminate infeasible solutions that lead to resource over-allocation. This mechanism ensures that suboptimal individuals are excluded subsequent generations, thereby avoiding the need to formulate a constrained linear optimization problem. A brief algorithmic overview of the GRANT method is presented in Algorithm 1 The crossover probability is set to a low value (0.1%), whereas the mutation is high, with a small mutation coefficient to maintain proximity to the near-optimal solution. While this configuration increases the risk of convergence to local minima, it significantly reduces computational complexity, making the approach suitable for real-time applications. As will be shown, GRANT achieves performance superior to existing methods.

Algorithm 1 GRANT Algorithm: individuals, representing possible allocations, are selected to minimize the sum of the users RAN delays.

Input: Population size (number of individuals) N_i, associated to an $M \times N$ allocation vector $\mathbf{b}$, number of generations N_G, crossover rate p_c, mutation rate p_m, selection method $\mathcal{S}$, fitness function of each individual $f(\mathbf{b}) = \sum_{0}^{N-1} \delta_n^{(RAN)}$.
Output: Best (minimum $f(\mathbf{b})$) solution found.
Initialize population $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{N_i}\}$
Evaluate the individual fitness $f(\mathbf{b_i})$, $i = 1, \dots N_i$
for generation = 1 **to** N_G **do**
  Select pairs of best fitting individuals
  for each pair $(\mathbf{b}_{i_1}, \mathbf{b}_{i_2})$, **do**
    if drawn r.v. $r < p_c$ **then**
      offspring -two new individuals- with their coefficients randomly selected half from $\mathbf{b}_{i_1}$ and $\mathbf{b}_{i_2}$ (crossover)
    else
      offspring individuals are ($\mathbf{b}_{i_1}$ and $\mathbf{b}_{i_2}$).
    end if
    for each offspring **do**
      if drawn r.v. $r < p_m$ **then**
        Update $\mathbf{b}_{i_k} = \mathbf{b}_{i_k} \odot \boldsymbol{\mu}$, where $\odot$ denotes the Hadamard product and μ is the vector the i.i.d. mutation coefficients $[\mu_0, \cdots \mu_{NM-1}]^T$ (mutation).
      end if
    end for
  end for
end for
Return: Best solution in the final population.

IV. Simulation Results

Herein, we consider a communication service scenario involving both mobile phone users and users of advanced XR services, such as Augmented Reality, Virtual Reality, and Control Gesture applications [19] in a setting where multiple base stations with 5G standalone connectivity are deployed. We examine a square area of side length $D_{Map} = 1000$m, within which M base stations (BSs) are positioned. Each BS has an equal maximum capacity constraint of $B_{max}^{(m)} = B_{max}$, $m = 0, \cdots M-1$, and a spectral efficiency that depends on the users' positions. Assuming closed-loop power control is implemented [20], we approximate the finite coverage of each BS as a circle of radius R. User services requests are generated in time slots of duration τs and are detected by a RAN Intelligent Controller. To reflect a realistic environment deployment scenario, users within the area are also associated with P edge servers, each with a capacity constraint $C_{max}^{(p)} = 5[GHz]$, $p = 0, \cdots P-1$.

The packet sizes are assumed to follow a truncated Gaussian distribution, as indicated in [19]. The packet play-out time is considered constant, and the through-

Variable	Description	Value
B_{max}	Maximum Bandwidth	$50[MHz]$
D_{Map}	Environment size	$100-1000[m]$
N	Number of Users	$30-1000$
M	Number of Base Stations	3
$\eta_n^{(m)}$	Spectral Efficiency	$25-35[bps/Hz]$
R	Base Station Transmission Radius	$[0.1-0.3]\cdot D_{Map}[m]$
L_{BS}	Max Linked BS	$3-5$

TABLE I: Numerical simulation parameters

put demand θ_n is modeled by a truncated Gaussian distribution, too. Users generate traffic associated with Augmented Reality (AR), Virtual Reality (VR), and Control Gesture (CG), and Mobile video services, with expected throughput Θ (Mbps) and Packet Delay Budget (ms) consistent with the specifications in in [19]. Specifically, the throughput Θ in [30-45 Mbps] for AR, [8-30 Mbps] for VR, and [10-30 Mbps] for CG, while the corresponding delay budget (ms) lies within [10-15 ms].

We evaluate the Quality of Experience (QoE) performance of the GRANT algorithm in terms of the packet delay budget, since in low-latency XR applications delays exceeding approximately 10ms can significantly degrad user-perceived quality. All the simulation parameters appear in Table I. The GRANT meta-parameters are set as follows: $N_G \leq N_G^{(max)} = 100$, $p_c = 0.1$, $p_m = 0.2$, $\Delta\mu = 0.2$, and $\mu_k,\ k=0,\cdots NM{-}1$ uniformly distributed in $1\pm\Delta\mu$.

We compare GRANT with different state of the art competitors. Specifically, we consider the cooperative greedy method in [10], which pursues XR-oriented Orchestration of Access Resources (X-OAR) by minimizing the overall RAN and MEC delay. Additionally, we consider the minimum delay allocation in [13], which operates in absence of cooperative scheduling, assuming that the n-th user is served by the closest base station that manages a bandwidth B_{TOT}. Therein, the optimal allocation of the users' bandwidths $B_n,\ \ n = 0,\cdots N-1$, in terms of minimum average delay at the RAN, is computed as: $B_n = \alpha\sqrt{\theta_n/(\eta_n\delta_n^{(MAX)})}$, where the factor α is selected such that $\sum_{n=0}^{N-1} B_n \leq B_{TOT}$. When the bandwidth required by a user to meet the delay constraints exceeds the available bandwidth, the allocated resource is proportional to the square root of the request. We refer to this allocation criterion as a Square Root Proportional (SRPROP) allocation. Finally, we assess the performance when the genetic algorithm is initialized with different conditions. For concreteness, we present the overall delay, which account for the RAN delay achieved by GRANT and the MEC delat from [10].

Let us now analyze the performance of the GRANT method in terms of delay and percentage of users whose packet delay budget has been met (satisfied users).

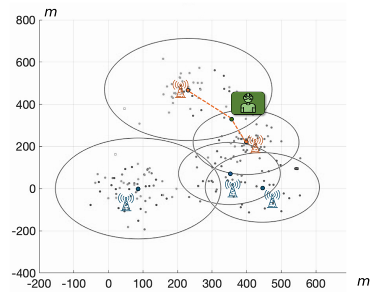


Fig. 2: Simulation topology: mobile and XR users served by RAN access points with overlapping coverages.

In Fig. 3(a), (b), and (c), we present the percentage of satisfied users versus the average delay for the GRANT method (red dot) across three different services: VR, AR, and CG. For comparison, we include results from the competitor algorithm in [10] (X-OAR, yellow dot) and the minimum-delay non-cooperative solution in [13] (SRPROP, green dot). Additionally, we evaluate a hybrid approach combining the X-OAR algorithm with a genetic-based approach (X-OAR+GA, blue dot). The GRANT method achieves lower delays while satisfying a higher percentage of users compared to both the greedy cooperative algorithm X-OAR and the optimal non-cooperative algorithm SRPROP.

The box plot in Fig. 4 illustrates the distribution of delays experienced by users under the GRANT method, categorized according to the type of service traffic. While the heuristic solution reallocates resources across stations to maximize satisfied users, GRANT explores solutions by adjusting allocation coefficients through a mutation process. This results in a broader delay distribution, highlighting the trade-off between increased computational complexity and reduced fairness among users within the same use-case, in exchange for a lower average delay.

As a final analysis, Fig. 5, presents the performance of the GRANT algorithm in terms of average delay

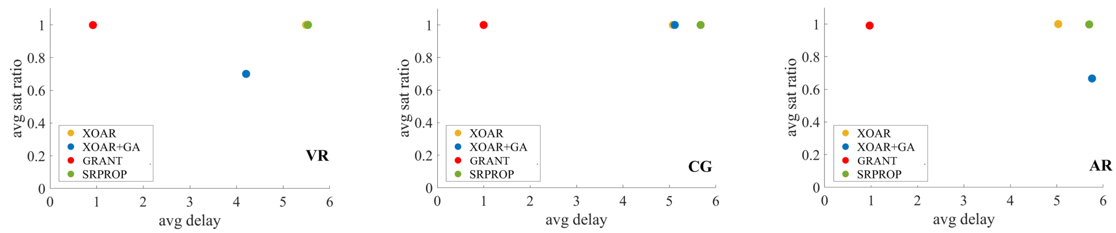


Fig. 3: Percentage of users whose delay requirement is satisfied versus the average delay for virtual reality services (a), control gesture services (b), and augmented reality services (c). The GRANT method (red dot) is consistently positioned in the top left of the plots, outperforming both cooperative (X-OAR, X-OAR+GA) and non-cooperative (SRPROP) competitors.

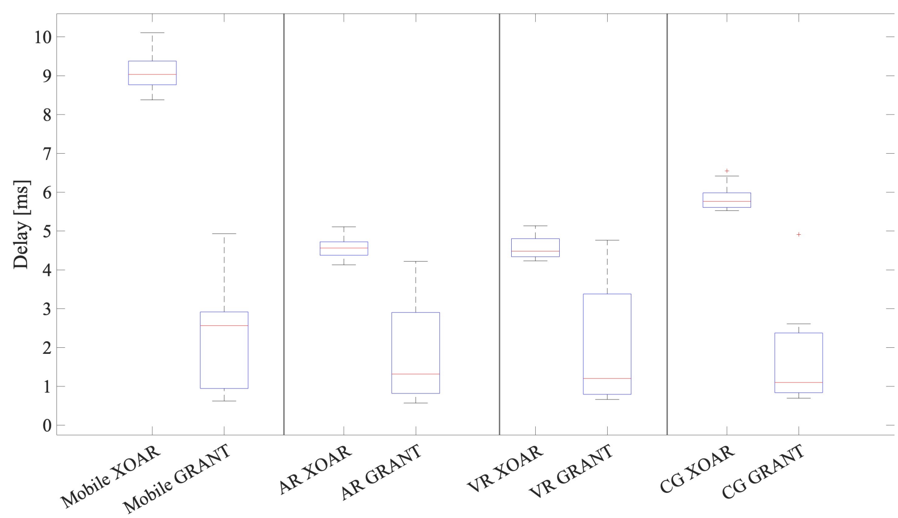


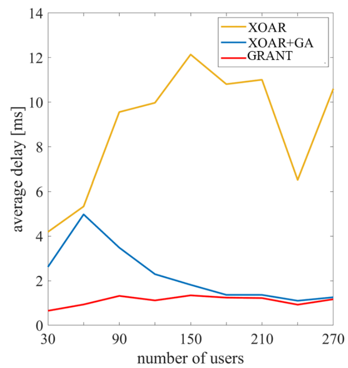
Fig. 4: Delay box plot the distribution of delays experienced by users with the GRANT method, divided according to the kind of service traffic.

and the percentage of satisfied users versus the number of users N. When resources are abundant, all allocation strategies maintain an acceptable number of satisfied users. However, as the number of users exceeds the threshold where resources become insufficient, the inherent randomness of the GRANT approach allows it to outperform the heuristic solution, which struggles to adapt effectively.

In summary, the allocation problem is solved by disjoint allocation followed by genetic optimization. Applying genetic algorithms on top of the X-OAR solution in [10] is suboptimal, as the X-OAR initialization traps the algorithm in a local minimum. In contrast, initializing bandwidth allocation with the SRPROP square root proportional distribution allows GRANT to converge within fewer than 20 generations, significantly outperforming state-of-the-art methods [10], [13].

GRANT achieves strong orchestration across varying spatial user densities. Among three initialization methods—minimum-delay disjoint allocation, greedy joint allocation, and genetic tracking—the minimum-delay disjoint allocation was most effective. Its computational complexity scales with the number of generations, population size, and fitness evaluation cost, with fitness evaluation as the most intensive step.

GRANT effectiveness is best understood through comparison with alternative strategies. When formulated with equality constraints on maximum delay under fully loaded base stations, GRANT aligns with the two-stage game-theoretic allocation in [14]: both begin with initial allocations meeting minimum user requirements, but GRANT applies constraints per base stations, whereas [14] applies constraints per user. Notably, GRANT incorporates cooperative scheduling for more complex and effective MAC-layer allocation. However, it does not yet consider application-layer rate adaptation, which remains unstandardized for XR data; extending cross-layer orchestration is a future direction.



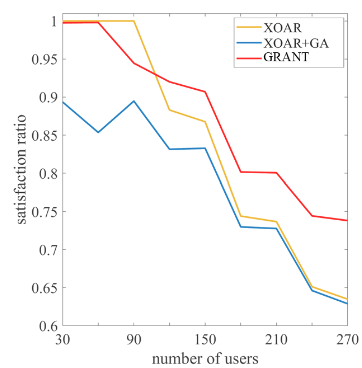


Fig. 5: Average delay (ms) and Percentage of satisfied users versus the number of users.

V. Conclusion and Future Work

In this paper, we presented GRANT, a novel genetic-based approach to RAN resource orchestration for extended reality verticals, targeting high-throughput, low-latency XR services. GRANT achieves minimum-delay radio access cooperative scheduling, enhancing XR user quality of experience compared to state-of-the-art solutions. It can be implemented as a controller of radio access resources, providing input to MAC-layer cooperative scheduling available in next-generation networks. Future work will extend resource orchestration by incorporating different QoS and QoE performance metrics into the GRANT cost function, as well as accounting for user traffic correlations, such as the spatial correlation of XR users in training applications.

References

[1] R. K. Horota, P. Rossa, A. Marques, L. Gonzaga, K. Senger, C. L. Cazarin, A. Spigolon, and M. R. Veronez, "An immersive virtual field experience structuring method for geoscience education," *IEEE Trans. on Learning Technologies*, vol. 16, no. 1, pp. 121–132, 2022.

[2] F. A. Fernandes, C. S. C. Rodrigues, E. N. Teixeira, and C. Werner, "Immersive learning frameworks: A systematic literature review," *IEEE Trans. on Learning Technologies*, 2023.

[3] M. Wang, H. Yu, Z. Bell, and X. Chu, "Constructing an edu-metaverse ecosystem: A new and innovative framework," *IEEE Trans. on Learning Technologies*, vol. 15, no. 6, pp. 685–696, 2022.

[4] M. Gapeyenko, V. Petrov, S. Paris, A. Marcano, and K. I. Pedersen, "Standardization of extended reality (xr) over 5g and 5g-advanced 3gpp new radio," *IEEE Network*, vol. 37, no. 4, pp. 22–28, 2023.

[5] I. F. Akyildiz and H. Guo, "Wireless communication research challenges for extended reality (xr)," *ITU Journal on Future and Evolving Technologies*, vol. 3, no. 1, pp. 1–15, 2022.

[6] F. Chai, Q. Zhang, H. Yao, X. Xin, R. Gao, and M. Guizani, "Joint multi-task offloading and resource allocation for mobile edge computing systems in satellite iot," *IEEE Trans. on Vehicular Technology*, 2023.

[7] P. Dai, M. Wu, K. Li, X. Wu, and Y. Ding, "Joint optimization for quality selection and resource allocation of live video streaming in internet of vehicles," *IEEE Trans. on Services Computing*, pp. 1–14, 2023.

[8] D. Zhang, L. Cao, H. Zhu, T. Zhang, J. Du, and K. Jiang, "Task offloading method of edge computing in internet of vehicles based on deep reinforcement learning," *Cluster Computing*, vol. 25, no. 2, pp. 1175–1187, 2022.

[9] M. Polese, M. Dohler, F. Dressler, M. Erol-Kantarci, R. Jana, R. Knopp, and T. Melodia, "Guest editorial open ran: A new paradigm for open, virtualized, programmable, and intelligent cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 241–244, 2024.

[10] A. Priviero, L. Mastrandrea, I. Chatzigiannakis, and S. Colonnese, "X-oar: Orchestration of access resources for extended reality educational applications," in *2024 IFIP Networking Conference (IFIP Networking)*, pp. 150–158, IEEE, 2024.

[11] S. Singh and D. H. Kim, "Joint optimization of computation offloading and resource allocation in c-ran with mobile edge computing using evolutionary algorithms," *IEEE Access*, vol. 11, pp. 112693–112705, 2023.

[12] X. Li, Z. Wang, Y. Sun, S. Zhou, Y. Xu, and G. Tan, "Genetic algorithm-based content distribution strategy for f-ran architectures," *ETRI Journal*, vol. 41, no. 3, pp. 348–357, 2019.

[13] S. Colonnese, F. Cuomo, T. Melodia, and I. Rubin, "A cross-layer bandwidth allocation scheme for http-based video streaming in lte cellular networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 386–389, 2016.

[14] S. Colonnese, F. Conti, G. Scarano, I. Rubin, and F. Cuomo, "Premium quality or guaranteed fluidity? client-transparent dash-aware bandwidth allocation at the radio access network," *Journal of Communications and Networks*, vol. 24, no. 1, pp. 59–67, 2022.

[15] S. D'Oro, L. Bonati, F. Restuccia, and T. Melodia, "Coordinated 5g network slicing: How constructive interference can boost network throughput," *IEEE/ACM Trans. on Networking*, vol. 29, no. 4, pp. 1881–1894, 2021.

[16] S. Karunarathna, S. Wijethilaka, P. Ranaweera, K. T. Hemachandra, T. Samarasinghe, and M. Liyanage, "The role of network slicing and edge computing in the metaverse realization," *IEEE Access*, vol. 11, pp. 25502–25530, 2023.

[17] Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Joint compute-caching-communication control for online data-intensive service delivery," *IEEE Trans. on Mobile Computing*, 2023.

[18] B. Gul, I. A. Khan, S. Mustafa, O. Khalid, S. S. Hussain, D. Dancey, and R. Nawaz, "Cpu and ram energy-based sla-aware workload consolidation techniques for clouds," *IEEE Access*, vol. 8, pp. 62990–63003, 2020.

[19] 3GPP, *Extended Reality (XR) in 5G, 3GPP TR 26.928*, 12 2020.

[20] S. Marinova and A. Leon-Garcia, "Intelligent o-ran beyond 5g: Architecture, use cases, challenges, and opportunities," *IEEE Access*, vol. 12, pp. 27088–27114, 2024.