IFIP Networking 2024 1571001667
2024 IFIP Networking Conference (IFIP Networking)

X-OAR: Orchestration of Access Resources for eXtended reality educational applications

Alessandro Priviero[1], Luca Mastrandrea[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*—Extended reality offers unprecedented learning and training occasions, and unique challenges related not only to throughput and delay, but also to the characteristic spatial concentration of trainees. We have developed an algorithm for eXtended reality oriented Orchestration of Access Resources (X-OAR) grounding on next generation network technologies. X-OAR is designed to efficiently allocate edge computing facilities and cooperatively scheduled radio access network resources for extended reality applications. Building on the 3GPP guidelines on quality of experience in XR services, X-OAR meets the stringent XR delay requirements by leveraging edge and radio resources and employing cooperative scheduling within the radio access network. We introduce a graph model of the X-OAR optimization problem, and we present the X-OAR greedy algorithm, that reduces the orchestration complexity and the dependency on user subscription information. Experimental results show that X-OAR, with its cooperative scheduling technique, outperforms state-of-the-art competitors in terms of XR quality of experience. X-OAR paves the way for further studies extending the system orchestration to the application layer and the related resource charging policy.**

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

I. Introduction

The proliferation of personal and mobile devices equipped with enhanced processing and networking capabilities has facilitated the realization of extended reality (XR) in various innovative domains. 5G technologies open new frontiers for immersive multimedia services by enabling shared experiences among geographically dispersed users [1]. At the same time the mobile edge computing (MEC) offered by 5G networks will allow mobile devices to rely on nearby processing facilities for low-latency real-time processing of video streams. By offloading computationally intensive tasks to nearby edge facilities, heat generation and battery drain on the mobile XR devices can be avoided and traded off with energy consumption at the resource side. Thus, the scarce energy at the user is preserved at the expenses of a heavier network energy load. Such technological advances enable the development of XR applications beyond the typical gaming and entertainment domains, towards the educational domain [2], [3]. By enabling the experience of remote situations [1], these technologies pave the way for effective training such as remote surgery or firefighting [4]. However, the effectiveness of immersive multimedia lies in high data throughput and minimal latency, which are crucial to ensure a satisfactory Quality of Experience (QoE) [5]. In some cases, XR education and training applications often require simultaneous, coordinated interaction of multiple users, relying on significant amount of low-latency traffic. This can lead to network congestion, especially when multiple users are simultaneously connected to a single base station or serviced by the same edge processing server at the same time [6]. In other cases, XR training scenarios are characterised by a high density of devices requiring significant radio access network (RAN) and/or edge processing capacity, for durations well beyond a few seconds (time slots) to hours. XR learning scenarios well fit ad-hoc designed 5G standalone infrastructures equipped with user charging points. The herein developed orchestration scheme tackles this case but straightforwardly generalizes to a wider spectrum of applications, where further energy or interference issues shall be accounted for.

Recently, several attempts at joint allocation of radio and computing resources have been proposed, usually involving different servers [7]–[9]. Existing solutions do not fully capitalize on the fact that the user may be covered by multiple radio access points. Radio transmission from nearby access points is typically considered as an interference, e.g., tackled by MIMO transmission design [10]. Recent research on vehicular networks with multiple access points with

This work was partially supported by the European Union's Horizon HADEA research and innovation program under grant Agreement 101092851 XR2LEARN project.

ISBN 978-3-903176-63-8 ©2024 IFIP

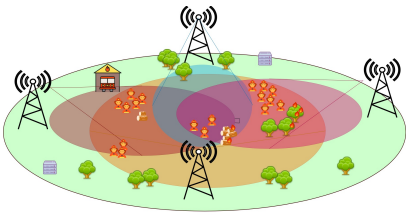

Fig. 1: XR training scenario: multiple base stations with overlapping coverage and edge servers located geographically close to each other allowing each user to connect to one or more access points and edge servers. X-OAR uses cooperative scheduling both in the RAN and at the edge.

overlapping radio coverage, has shown advantages in terms of resource allocation [11]–[13]. Herein, we exploit the multiple base stations coverage by cooperative radio resource scheduling available in 5G and beyond networks [14], [15]. We also assume that users can offload different tasks, such as rendering different views with overlay information, to one or more geographically accessible servers [16], [17].

We present here the e**X**tended reality oriented **O**rchestration of **A**ccess **R**esources (X-OAR) method for XR training groups of users that may fall under the overlapping coverage of different Radio Access Network (RAN) base stations, and may be provided with shared Mobile Edge Computing (MEC) resources. The X-OAR method pursues throughput guarantees and delay minimisation at the application layer leveraging cooperative capabilities offered by 5G network function virtualization both at the RAN and at the edge [14], [15], [18], [19].

To derive X-OAR, we first formulate the problem with target delay and throughput requirements, assuming cooperative RAN scheduling at different base stations and MEC scheduling at different servers. We then derive X-OAR as a greedy allocation policy that induces users to use resources from multiple suppliers (either base stations or servers), addressing radio access and edge computing server resource orchestration. The main novel contributions of the paper are as follows:

- X-OAR explicitly targets a spatially dense, low-latency XR user environment;
- X-OAR uses radio cooperative scheduling between different base stations, in parallel with multi-server resource orchestration.

Numerical simulations validate that X-OAR outperforms state-of-the-art solutions in 3GPP-compliant realistic scenarios by exploiting the flexibility of next-generation networks.

The paper is organised as follows: after reviewing the XR training services in Section II, we outline the orchestration policy in Section III and present the X-OAR method in detail in Section IV. Numerical simulations in Section V demonstrate the superiority of X-OAR over current solutions. Sec. VI concludes the paper and discusses future research directions.

II. XR Enhanced Education Service Architecture

We consider XR education and training scenarios where one or more groups of people operate mobile XR devices and interact within and across groups. The network traffic and processing capacity requirements of each user are different due to the individual devices and the personalised XR application being run. The interactions between users have strict delay requirements [20] and require increased throughput in both up-link and down-link. In such scenarios, the orchestration goal is to deliver high quality services to XR users that are training under different critical scenarios [21].

While processing and transmitting immersive video streams for education and training, video QoE indexing techniques, such as buffer occupancy or estimated throughput, cannot be used. Due to their interactive, low-delay nature, XR services have delay requirements so stringent that the transmission of lost packets is avoided [22]. Table I summarises the expected throughput Θ (Mbps) and Packet Delay Budget (PDB) (ms) for Augmented Reality (AR), Virtual Reality (VR), and Control Gesture (CG) for XR services [22]. For comparison sake, the throughput and PDB of mobile voice traffic services are also shown. In the following, we characterise the QoE of XR service in terms of user delay and throughput; further quality features tailored to the user profile and service content are currently under investigation [23].

TABLE I: Expected throughput Θ (Mbps) and Packet Delay Budget (PDB) (ms) for Augmented Reality (AR), Virtual Reality (VR), Control Gesture (CG) XR services and for Mobile video services (source: [22]).

	AR	VR	CG	Mobile video
Θ (Mbps)	30-45	30-45	8-30	10-30
PDB (ms)	10	10	15	40

Given such challenging QoE constraints, the development of specific resource allocation strategies is required. Many efforts have been proposed trying to address resource allocation in networks with stringent requirements, including digital twins of networks, as described in [24], which provide valuable insight into resource allocation and end-to-end quality metric calculation. The relevant literature addresses RAN

resource allocation and data throughput on the XR application layer [25], [26], while the impact of MEC resources on XR frame rendering and processing offloading is an open research topic [27].

We consider XR education/training scenarios where the users are remotely connected to a control room that streams data to the training scene based on scenario-specific criteria that can change dynamically as an education/training session evolves. The training area can be connected by multiple base stations, with 5G standalone connectivity. Unlike traditional approaches that rely on single base station coverage for XR service provisioning, X-OAR introduces a novel graph-based model and greedy algorithm that significantly improves the quality of experience by taking full advantage of 3GPP and 5G New Radio flexible resource allocation, where the multi-cell MAC/PHY layer constraints (bandguard, resource block allocation distance) can be cooperatively managed on a milliseconds time scale, as in [15].

X-OAR addresses the cooperative scheduling architecture at a higher functional level, and the X-OAR solution can drive the lower level cooperative scheduling [15], which operates at a finer temporal scale and takes into account radio interference or overlapping of allocated resources, e.g. when two users share the same resource block at the intersection of the two base stations serving them.

III. XR-oriented Orchestration of Access Resources (X-OAR): Methodology

The objective of X-OAR is to optimize user QoE by orchestrating RAN and MEC resources. The main quality parameters considered here with respect to the XR application are data throughput and temporal XR fluidity, which are directly related to the packet delay at the application layer. The data throughput is usually determined at the application layer [28], and it is assumed here as an input constraint. The delays generated by the allocated RAN and MEC resources both contribute to the delay experienced by the user. Therefore, in X-OAR we adopt a symmetric model of the RAN and MEC sides of the allocation problem, as shown in Fig. 2. This model, which takes into account the actual coverage of multiple base stations' and the availability of edge computing servers, addresses the scenario of XR training services, where a large number of XR users with high throughput requirements are located in high-density areas.

The orchestration problem can be abstracted by a graph whose nodes represent the base stations, the edge servers and the users while the links represent an assignment of a user to a base station and an edge server. The links are characterised using i) spectral efficiency for the radio side, since the channel quality

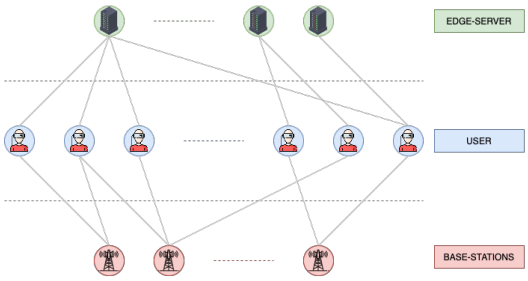

Fig. 2: Graph model of the correspondences between users and the different resources.

metric affects the throughput achieved by the amount of bandwidth allocated, and ii) a computational efficiency for the edge resources, to characterise the effectiveness of the offload. In the following, let N be the number of users, M the number of base stations and P the number of edge computing servers, both managed by an integrated manager (IM), e.g. the infrastructure provider in [15], [24], [29].

We assume that the execution of an XR application is divided into time slots of equal length. In the τ-th time slot, the n-th user transmits/receives packets of different traffic types, i.e., AR, VR, CG and Mobile video. Let $\theta_n[i]$ denote the throughput requested by the n-th user at the τ-th time slot, including audio data and control gesture data. Let $\mathcal{W}(\theta_n[\tau])$ denote the computational load, depending on the complexity, of the XR content. We assume here that computation load is connected to throughput, and introduce the weight ρ such that $\mathcal{W}(\theta_n[\tau]) = \rho \cdot \theta_n[\tau]$. Without loss of generality, in the sequel we consider the allocation for a single time slot, thus for simplicity dropping τ.

The n-th user quality profile Q_n is characterized by the minimum requested throughput $\theta_n^{(min)}$ and maximum tolerated delay $\delta_n^{(max)}$, i.e., $Q_n = \mathcal{Q}\left(\theta_n^{(min)}, \delta_n^{(max)}\right)$. Assuming that the throughput θ_n is selected at the application layer so as to meet the requirement $\theta_n \geq \theta_n^{(min)}$, we focus on the delay of the n-th user and express it as the sum of the RAN and MEC delays, both related to the n-th user required throughput θ_n; specifically, the RAN delay θ_n and the MEC delay δ_n contribute to a fraction t_n and $(1 - t_n)$ of the total delay δ_{tot}, respectively.

Depending on their locations, the users are covered by one or more base stations, with different spectral efficiencies; we denote the spectral efficiency between the m-th base station and the n-th user by $\eta_n^{(m)}$, where $\eta_n^{(m)}$ vanishes in the limit case when the user is outside the coverage of the base station. Furthermore, each user has potential access to a subset of servers depending on their subscription profile. We charac-

terise the server accessibility and the processing and rendering capabilities [30] by a scalar computational efficiency parameter. The computational efficiency between the p-th edge server and the n-th user is denoted by $\beta_n^{(p)}$, which is zero in the limit case when the user has no subscription access to the server.

The RAN bandwidth allocated by the m-th base station to the n-th user is denoted as B_n^m and the allocated MEC computational capacity of the p-th edge server as C_n^p so as to meet the user quality profile Q_n. The delay is formulated as follows:

$$\delta_n = \frac{\theta_n}{\sum_{m=0}^{M-1} B_n^{(m)} \cdot \eta_n^{(m)}} + \frac{\mathcal{W}(\theta_n)}{\sum_{p=0}^{P-1} C_n^{(p)} \cdot \beta_n^{(m)}} \quad (1)$$

From the above equation we can see that the n-th user delay δ_n jointly depends on the effective RAN throughput

$$\mathcal{B}_n = \sum_{m=0}^{M-1} \eta_n^{(m)} \cdot B_n^{(m)}$$

and the effective MEC computational effort

$$\mathcal{C}_n = \sum_{p=0}^{P-1} \beta_n^{(p)} \cdot C_n^{(p)}$$

We aim to optimize the RAN and MEC resource allocation:

$$\left(B_n^{(m)}, \cdots, C_n^{(p)} \cdots\right)$$

so as to minimize the overall delay

$$\min_{B_n^{(m)} \cdots C_n^{(p)}} \sum_{n=0}^{N-1} \delta_n \left(\mathcal{B}_n, \mathcal{C}_n\right) \quad (2)$$

under the constraints over the actual base stations' bandwidths and servers' capacities:

$$\begin{aligned} &\sum_{n=0}^{N-1} B_n^{(m)} \leq B_{TOT}^{(m)},\ m \in \{0, \ldots, M-1\} \\ &\sum_{n=0}^{N-1} C_n^{(p)} \leq C_{TOT}^{(p)},\ p \in \{0, \ldots, P-1\} \end{aligned} \quad (3)$$

For a given user's throughput request θ_n the delay δ_n is obtained by a family of pairs of effective bandwidth and capacity values $(\mathcal{B}_n, \mathcal{C}_n)$. Fig.3 depicts different examples of the delays versus the effective bandwidth and capacity $\mathcal{B}, \mathcal{C}$ allocated for different users: the yellow points in the plots represent the set of pairs $(\mathcal{B}_n, \mathcal{C}_n)$, $n = 0, \cdots, N-1$ that should be jointly optimized. It is therefore up to the Integrated Manager (IM), responsible for the orchestration [29], to select an optimal set of pairs $(\mathcal{B}_n, \mathcal{C}_n), \forall\, n \in [0, N-1]$.

IV. X-OAR Greedy Algorithm

An optimal cooperative resource allocation requires a global knowledge of the QoE requirements of the users along with their locations. Since each service provider has its own physical infrastructure and subscription policy, e.g., providing the user with access to different resources, the integrated manager may not have contemporary and updated access to such information.

In this section we introduce the X-OAR greedy solution that handles the allocation of the RAN and MEC separately by a two-tier algorithm as illustrated in Fig. 3. The first stage initializes the allocation by separately accounting for the quality constraint and identifying a Minimum Allocated Resources (MAR) point for each user. Then, if the MAR solution violates either a resource or a quality constraint, a selected set of heavily demanding users undergoes a resource adaptation till the constraints are met.

A. Initialization

At the initialization step, the starting set of the Minimum Allocated Resources (MAR) point for each user is calculated by symmetrically distributing the delay contributions on the RAN and MEC stages, i.e.,

$$\frac{\mathcal{B}_n^{(\text{MAR})}}{\theta_n} = \frac{\mathcal{C}_n^{(\text{MAR})}}{\theta_n} = \frac{\delta_n^{(max)}}{2}$$

for $n \in \{0, \ldots, N-1\}$. Fig.3 depicts the delay δ_n versus the effective computational capacity $\mathcal{C}_n$ and bandwidth $\mathcal{B}_n$ for different users. The delay constraint implied by the user quality profile $Q_n = \left(\theta_n, \delta_n^{(max)}\right)$ is identified by an iso-level curve of the plane $(\mathcal{B}_n, \mathcal{C}_n)$ corresponding to the maximum tolerable delay value. The pair $\left(\mathcal{B}_n^{(\text{MAR})}, \mathcal{C}_n^{(\text{MAR})}\right)$ is the intersection of the first quadrant bisector with the delay level curve at $\delta_n^{(max)}/2$.

Based on the users' and base stations positions we construct a matrix of spectral efficiencies. We represent the coverage by the $M \times N$ matrix A_{RAN}:

$$A_{\text{RAN}} = \begin{bmatrix} \eta_0^{(0)} & \cdots & \eta_0^{(M-1)} \\ \vdots & \ddots & \vdots \\ \eta_{N-1}^{(0)} & \cdots & \eta_{N-1}^{(M-1)} \end{bmatrix} \quad (4)$$

whose (m, n)-th element equals the spectral efficiency $\eta_n^{(m)}$ of the channel between the n-th user and the m-th base station. The same approach can be used for the computational efficiencies, we construct the $P \times N$ matrix A_{MEC}:

$$A_{\text{MEC}} = \begin{bmatrix} \beta_0^{(0)} & \cdots & \beta_0^{(P-1)} \\ \vdots & \ddots & \vdots \\ \beta_{N-1}^{(0)} & \cdots & \beta_{N-1}^{(P-1)} \end{bmatrix} \quad (5)$$

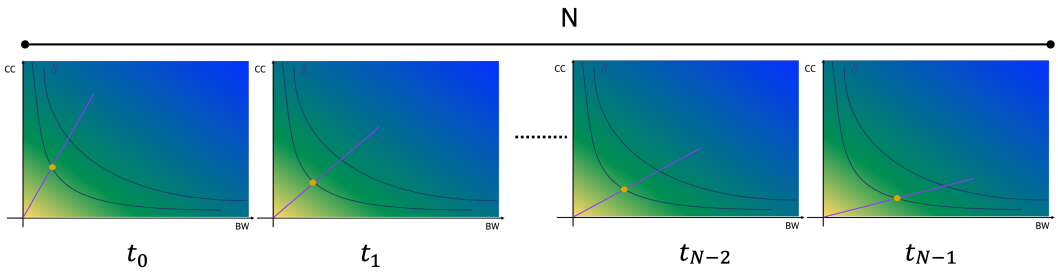

Fig. 3: Delay δ_n versus effective computational capacity $\mathcal{C}_n$ and bandwidth $\mathcal{B}_n$ for different users. X-OAR solution of (2) corresponds to a set of $(\mathcal{B}_n, \mathcal{C}_n)$ pairs (yellow points); the X-OAR greedy method firstly equates the delays of the RAN and MEC stages and then adapt them till system constraints are met.

whose (p, n)-th element represent the computational efficiency $\beta_n^{(p)}$ of the p-th server with respect to the n-th user.

Thanks to the symmetric problem structure, once the target effective RAN throughput and MEC computational effort, namely $\mathcal{B}_n$, $\mathcal{C}_n$, for $n \in \{0, \dots N-1\}$, are set, the optimization problem decouples as follows. Let us introduce the vectors collecting the unknown target allocations:

$$\mathbf{b} = \left[B_0^{(0)}, \cdots, B_{N-1}^{(0)}, \cdots, B_0^{(M-1)}, \cdots, B_{N-1}^{(M-1)}\right]^T, \tag{6}$$

$$\mathbf{c} = \left[C_0^{(0)}, \dots, C_{N-1}^{(0)}, \dots, C_0^{(P-1)}, \dots, C_{N-1}^{(P-1)}\right]^T, \tag{7}$$

of size $MN \times 1$ and $PN \times 1$, respectively. Then, we formulate the bandwidth resource allocation problem as the following linear system, where $\odot$ denotes the Khatri-Rao (column-wise Kronecker) product:

$$(A_{RAN} \odot I_N)^T \cdot \mathbf{b} = \begin{bmatrix} \mathcal{B}_0^{(\text{MAR})} \\ \vdots \\ \mathcal{B}_{N-1}^{(\text{MAR})} \end{bmatrix} \tag{8}$$

under the constraints on the base stations capacity

$$(I_M \odot \mathbf{1}_N) \cdot \mathbf{b} = \begin{bmatrix} B_{\text{TOT}}^{(0)} \\ \vdots \\ B_{\text{TOT}}^{(M-1)} \end{bmatrix} \tag{9}$$

The same linear problem applies to the computational capacity allocation:

$$(A_{MEC} \odot I_N)^T \cdot \mathbf{c} = \begin{bmatrix} \mathcal{C}_0^{(\text{MAR})} \\ \vdots \\ \mathcal{C}_{N-1}^{(\text{MAR})} \end{bmatrix} \tag{10}$$

under the constraint on the edge server's capacity

$$(I_P \odot \mathbf{1}_N) \cdot \mathbf{c} = \begin{bmatrix} C_{\text{TOT}}^{(0)} \\ \vdots \\ C_{\text{TOT}}^{(P-1)} \end{bmatrix} \tag{11}$$

B. Adaptation

After the initialization phase, it may occur that the solution based on the MAR set of pairs, obtained by requiring an equal balance between the RAN and MEC delays, violates certain capacity constraint. This may occur, for instance, when a user undergoes a high competition by neighbouring users on the RAN bandwidth resources, while the associated servers are not fully loaded. In this case, we improve the overall system performance by changing the delay contribution balance to relax the RAN delay requirement while restricting the MEC delay requirement. This is illustrated in Fig.3, where we recognize that, instead of assuming $t_n = 0.5$, the two components of the delay (1) can be balanced differently. For the n-th user, the pair $(\mathcal{B}_n, \mathcal{C}_n)$ is associated to a trade-off parameter t_n that represents the balance between the delay contribution of each resource, namely

$$\mathcal{B}_n = t_n \frac{\delta_{tot}}{\theta_n}, \qquad \mathcal{C}_n = (1 - t_n) \frac{\delta_{tot}}{\theta_n}$$

the trade-off parameter t_n controls the normalized slope of the straight line that connects the origin of the axis $(0, 0)$ and the pair in Fig.3. In this light, the optimization solution associated to N pairs $(\mathcal{B}_n, \mathcal{C}_n)$, actually corresponds to identifying an optimal vector of trade-off parameters $\mathbf{t} = \{t_0, t_1, \dots, t_{N-1}\}$ so as to meet the problem constraints.

The greedy X-OAR stems from this observation and can be formulated as follows: after solving the linear problem for equal transmission and computing delays, i.e., the MAR solution $t_n = 0.5$ for all the users, the allocation is adapted in the following steps:

S1 checking if the constraints

$$B^{(m)} \leq B_{\text{TOT}}^{(m)}, C^{(p)} \leq C_{\text{TOT}}^{(p)}$$

are met for all m and p;

S2 if the constraints are violated on the m-th base station or on the p-th server, selecting the most resource demanding (heavy) user;

S3 transferring the heavy users' load to other available resources and
S4 sharing the released bandwidth and capacity among all the active users.

In more detail, for the i-th overloaded resource (either a base station or a server) let the j-th users with the highest requested throughput θ_j be selected. Then, the t_j value is adjusted to reduce the demand on the overloaded resource. If there is no available complementary resource, e.g. MEC resource in the case of RAN overload, to address the constraint, users are dropped based on their requested throughput θ, starting with the highest, until the constraint is met. Dropping all the heavy users will avoid overload but also release resources. The released resources are distributed equally among connected users to improve their perceived quality.

C. Further considerations

The X-OAR greedy algorithm disregards any application layer rate, because while buffer-based or throughput rate adaptation plays an important role in the development of mobile streaming services [31], the development of XR-appropriate adaptation is still an open research question [32].

Furthermore, we assume that each user can, in principle, be connected to all covering base stations; however, this comes at the cost of increased signaling and user profile sharing, leading to a trade-off between coverage optimisation and orchestration complexity and flexibility. We address this trade-off by setting a maximum number of base stations to which users can be connected. For the same reason, we also limit the number of servers to which each user is connected. In addition, the computational efficiency weight $\mathcal{W}(\theta_n)$ highlights the dependency of the delay on the amount of data offloaded; thus, it includes both the offload latency due to transmission and the computational latency [20].

Lastly, the total energy consumed at the user's device includes the energy consumed at the device for computation, transmission and offloading, as well as for the idle state. It is assumed that the XR data is processed at the edge in the downlink, for example for network assisted rendering. If the MEC server is co-located with the base station the computation task is performed during data transmission from the base station to the network, and from the mobile user's point of view there is no additional transmission energy consumption. However, the energy consumption in the network increases; the detailed analysis of the energy cost in the network depends on the geographical location of the edge servers. In the common case where each base station includes edge computing server facilities, the in-network energy transmission is virtually negligible. Otherwise, the energy cost depends on the actual geographical server deployments. We leave a full analysis of energy consumption to future studies.

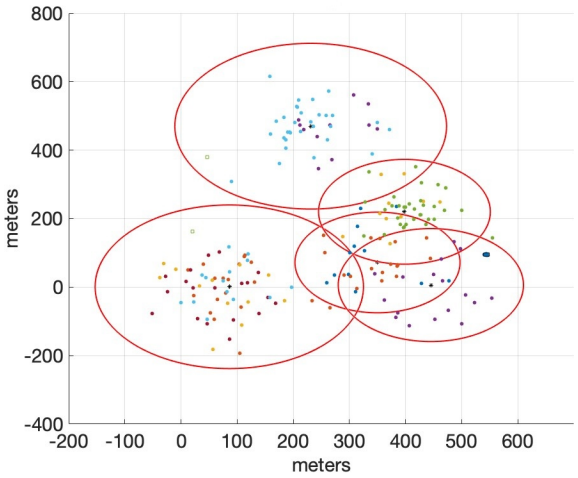

Fig. 4: Simulation topology Example (Firefighters case): clusters of XR training users are concentrated in an area served by different RAN access points

V. Experimental evaluation

Variable	Description	Value
B_{max}	Maximum Bandwidth	$200[MHz]$
C_{max}	Maximum Computational Capacity	$5[GHz]$
D_{Map}	Environment size	$100-1000[m]$
N	Number of Users	$30-1000$
M	Number of Base Stations	$3-15$
P	Number of Edge Servers	$2-5$
$\eta_n^{(m)}$	Spectral Efficiency	$25-35[bps/Hz]$
$\beta_n^{(p)}$	Computational Efficiency	$3-5[bps/Hz]$
ρ	Computational weight	$\rho = 0.8$
R	Base Station Transmission Radius	$[0.1-0.3] \cdot D_{Map}[m]$
$t^{(0)}$	Initial Trade-off	0.5
L_{BS}	Max Linked BS	$3-5$
N_{MC}	Montecarlo runs	10

TABLE II: Experimental evaluation parameters

We validate the effectiveness of X-OAR by numerical experiments that replicate various scenarios. We compare X-OAR with the proportional algorithm (Prop) in [33], which allocates the bandwidth achieving the minimum average user delay for a given BS bandwidth amount, without exploiting cooperative scheduling. We demonstrate that X-OAR achieves better performance in terms of application layer packet delay and percentage of users served thanks

to its flexible, cooperative, resource orchestration strategy.

We considered two characteristic scenarios: the first scenario involves the training of future firefighters [4]. The new recruits are trained to scenarios that often cannot be replicated, so XR can be used to simulate challenging situations. We considered a scenario where a control room could control sensors in the training camp and visualise the entire scene. The members of the control room and the trainees in the training camp are equipped with XR devices. The control room and the outdoor training camp are connected by multiple base stations with 5G standalone connectivity.

For the second scenario, we considered different groups of medical students simulating a remote surgery. Each group of students is placed in a different room, the operating room is equipped with sensors and cameras and a number of participating nurses are also equipped with XR equipment. The students use the XR equipment to carry out the surgical operation, and track in high precision the movement of the hands. This is a worst case scenario, where the XR training takes place in a restricted area under the coverage of only one B, and cooperative scheduling is hindered.

In both the scenarios, we consider a square area of side $D_{Map} = 1000$m where M base stations (BSs) are located; the users in the area are also associated to P edge servers (ESs). The BSs have a finite coverage given by a radius R, with equal maximum capacity constraint of $B^{(m)}_{max} = B_{max}$, $m = 0, \cdots M-1$, and a spectral efficiency depending on the users' positions. The ESs have maximum capacity constraint $C^{(p)}_{max} = C_{max}$, $p = 0, \cdots P-1$ and an efficiency ρ. The user requests are generated in time slots of τs, and they are detected by an Integrated Manager (IM) that implements the greedy algorithm of Sec.IV. The algorithm modifies the trade-off parameter t_n of each user, checking if there is any user that needs to be dropped out. All the simulation parameters can be found in Table II.

The packet sizes are distributed according to a truncated Gaussian distribution as in [22]. The packet play-out time is constant, and θ_n follows a truncated Gaussian distribution, too. The users generate AR, VR, CG and Mobile video traffic characterized by different constraints as in Table I.

We evaluate the QoE performance of the X-OAR greedy algorithm by following the 3GPP guidelines, where the user QoE is quantified in terms of target packet delay budget and throughput. This reflects the XR application need for low latency: packets delays which exceed a certain threshold- about 10 ms according to 3GPP guidelines- heavily affect the perceived quality of the user. The X-OAR greedy method addresses both RAN and MEC allocation with assigned QoE constraint, by leveraging cooperative RAN scheduling.

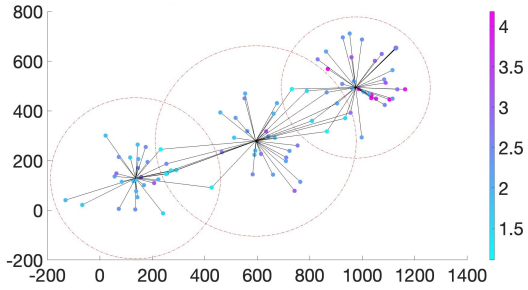

Fig. 5: Weighted Topology: the colors of the points represent the user delays and the users-to-base-stations connection.

We provide a visual example of X-OAR principles and performance in Fig. 5, obtained using the parameters in Tab. II. Each point represents a user, and its color represents the experienced delay: we recognize that users covered and cooperatively served by different base stations achieve smaller delays than users covered by a single base station. The effect is more visible on heavy loaded base stations (upper right part of the plot) than on less crowded areas (lower left part of the plot).

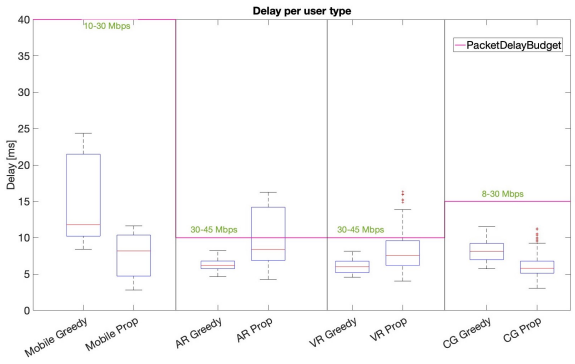

Fig. 6: Boxplots of average values and quartiles of the delays observed on different class of traffics for the greedy X-OAR (Greedy) and proportional (Prop) [33] algorithm. The red horizontal bars represent the maximum delay per class

In the simulations, we set the maximum number of servers to which a user can be subscribed to. For the time span of the simulation, each user is randomly linked to a number of servers less or equal to this maximum. From a technical standpoint, the C_{max} parameter is the same for all servers at 5GHz (assuming a Xeon CPU), and the computational efficiencies,

ranging from 3 to 5, are fixed for the duration of the simulation.

The user resources are allocated using the X-OAR method. For each kind of traffic, we collect the users' delays over $N_{MC} = 10$ Montecarlo runs. The boxplots in Fig 6 show the average values and the quartiles of the delays observed for different class of traffics using the greedy X-OAR algorithm (Greedy) and the suitable proportional algorithm (Prop) in [33]. The red horizontal bars represent the maximum tolerated delay per each class, and the related mean per user throughput is also reported: we recognize that the strictest delay requirements are met in the AR and VR classes. The Mobile Class is more delay tolerant. Using X-OAR, the packet delay budget is achieved for all the three classes; noticeably, even the AR and VR heavier traffic classes face a delay below the budget, whereas the proportional algorithm in [33] fails to meet these constraints. The reason why X-OAR succeeds is that it allows a larger delay on the Mobile traffic class, fully exploiting the looser packet delay budget conditions.

Fig. 7 provides further insight on X-OAR action, plotting: the average per user delay (left axis) and the percentage of users that meet the packet delay budget -satisfied users- (right axis) as a function of the maximum number of covering base stations to which a user can connect. In situations where resources are limited, X-OAR leverages cooperative scheduling and having more stations available for connection significantly enhances user satisfaction and reduces average delay. Even with few (e.g. 2) base stations the X-OAR Greedy algorithm satisfies more than 80% of the users, whereas the allocation in [33] remains under 60% of satisfied users. Fig. 7 compares the performances in different conditions. The MedSchool scenario hinders cooperative scheduling and its performance are observed for a number of cooperative base stations equal to one. The delay of the Greedy algorithm (blue solid line) is half of the delay achieved by the allocation in [33] (proportional, red solid line). The number of users satisfied by the Greedy algorithm (blue dashed line) is also higher than that of proportional (red dashed line). When a larger number of BS is adopted, as in the FireFighters scenario, the overall bandwidth increases; X-OAR exploits the cooperative scheduling systematically outperforms the method in [33], definitely increasing the number of satisfied user.

Tab. III reports the fairness of X-OAR over different traffic classes. X-OAR guarantees the fairness, which remains above 0.9 over all the traffic classes, and in most cases it outperforms the competitor.

Finally, we observe that the performance of the X-OAR algorithm depends the overall bandwidth and

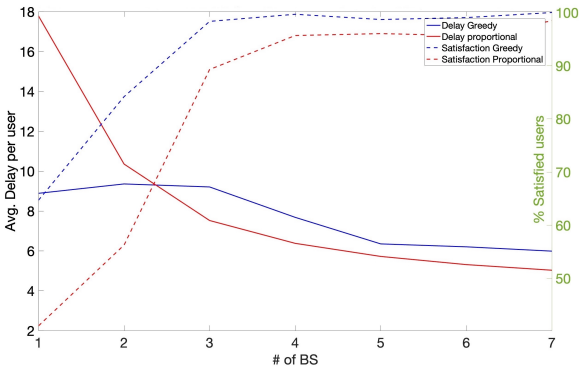

Fig. 7: Average delay (left axis) and percentage of satisfied users (right axis) versus the maximum number of cooperative base stations per user for X-OAR (blu lines) and [33] (red lines).

TABLE III: Fairness of X-OAR Greedy algorithm for different traffic classes.

Use Case	Greedy	Proportional
Mobile video	0.9130	**0.9415**
AR	**0.9643**	0.9580
VR	**0.9635**	0.9565
CG	**0.9506**	0.9279

the number of users. Fig.8 plots the average delay as a function of the number of users, i.e. of the RAN load. The X-OAR delay is stable, and the advantage over conventional methods increases with the load.

VI. Conclusion and Future Work

In this paper, we presented a novel approach to resource orchestration for high throughput, low latency, XR educational and training services, which often risk causing access network overload due to the spatial concentration of users. Our novel method, X-OAR, combines radio access cooperative scheduling and computational capacity allocation, enhancing XR user quality of experience compared to state-of-the-art solutions. X-OAR resorts to an Integrated Manager for intelligent resource allocation, fully

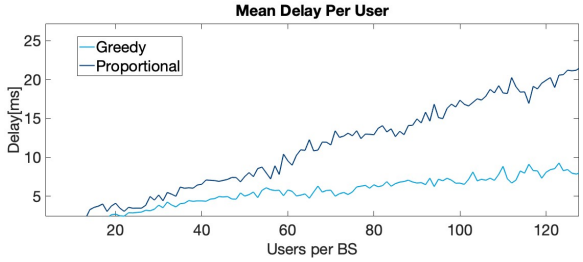

Fig. 8: Average delay, over all the users and the traffic classes, versus the number of users

leveraging cooperative scheduling available in next-generation networks. X-OAR adopts a graph based problem representation, setting the stage for a graph neural network-based, data-driven implementation. Future work will extend resource orchestration to cooperative XR rate adaptation and resource charging policy, laying the groundwork for a Digital Twin of XR training frameworks.

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] S. G. Wheeler, H. Engelbrecht, and S. Hoermann, "Human factors research in immersive virtual reality firefighter training: A systematic review," *Frontiers in Virtual Reality*, vol. 2, p. 671664, 2021.

[5] 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.

[6] 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.

[7] Y. Zhou, F. R. Yu, J. Chen, and B. He, "Joint resource allocation for ultra-reliable and low-latency radio access networks with edge computing," *IEEE Trans. on Wireless Communications*, vol. 21, no. 1, pp. 444–460, 2021.

[8] L. Ferdouse, A. Anpalagan, and S. Erkucuk, "Joint communication and computing resource allocation in 5g cloud radio access networks," *IEEE Trans. on Vehicular Technology*, vol. 68, no. 9, pp. 9122–9135, 2019.

[9] A. Garcia-Saavedra, G. Iosifidis, X. Costa-Perez, and D. J. Leith, "Joint optimization of edge computing architectures and radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2433–2443, 2018.

[10] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[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] L. Huang, X. Feng, L. Zhang, L. Qian, and Y. Wu, "Multi-server multi-user multi-task computation offloading for mobile edge computing networks," *Sensors*, vol. 19, no. 6, 2019.

[17] W. Zhang, Y. Zhang, Q. Wu, and K. Peng, "Mobility-enabled edge server selection for multi-user composite services," *Future Internet*, vol. 11, no. 9, 2019.

[18] 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.

[19] 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.

[20] A. Medeiros, A. Di Maio, T. Braun, and A. Neto, "Service chaining graph: Latency-and energy-aware mobile vr deployment over mec infrastructures," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 6133–6138, IEEE, 2022.

[21] J. C. Uhl, G. Regal, H. Schrom-Feiertag, M. Murtinger, and M. Tscheligi, "Xr for first responders: Concepts, challenges and future potential of immersive training," in *International Conference on Virtual Reality and Mixed Reality*, pp. 192–200, Springer, 2023.

[22] 3GPP, *Extended Reality (XR) in 5G, 3GPP TR 26.928*, 12 2020.

[23] P. Panarese, A. Baiocchi, and S. Colonnese, "The extended reality quality riddle: A technological and sociological survey," in *2023 International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 1–6, IEEE, 2023.

[24] H. Wang, Y. Wu, G. Min, and W. Miao, "A graph neural network-based digital twin for network slicing management," *IEEE Trans. on Industrial Informatics*, vol. 18, no. 2, pp. 1367–1376, 2020.

[25] D. Brunori, S. Colonnese, F. Cuomo, G. Flore, and L. Iocchi, "Delivering resources for augmented reality by uavs: a reinforcement learning approach," *Frontiers in Communications and Networks*, vol. 2, p. 709265, 2021.

[26] K. Koutlia, B. Bojovic, S. Lagén, X. Zhang, P. Wang, and J. Liu, "System analysis of qos schedulers for xr traffic in 5g nr," *Simulation Modelling Practice and Theory*, vol. 125, p. 102745, 2023.

[27] D. G. Morin, D. Medda, A. Iossifides, P. Chatzimisios, A. G. Armada, A. Villegas, and P. Perez, "An extended reality offloading ip traffic dataset and models," 2023.

[28] R. Doreswamy, A. G. Colaco, V. Sevani, P. Patil, and H. Tyagi, "Rate adaptation for low latency real-time video streaming," in *2023 National Conference on Communications (NCC)*, pp. 1–6, IEEE, 2023.

[29] 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.

[30] D. Yuan, E. Hossain, D. Wu, X. Liu, and G. Dudek, "Realizing xr applications using 5g-based 3d holographic communication and mobile edge computing," 2023.

[31] S. Colonnese, P. Frossard, S. Rinauro, L. Rossi, and G. Scarano, "Joint source and sending rate modeling in adaptive video streaming," *Signal Processing: Image Communication*, vol. 28, no. 5, pp. 403–416, 2013.

[32] B. Bojović, S. Lagén, K. Koutlia, X. Zhang, P. Wang, and L. Yu, "Enhancing 5g qos management for xr traffic through xr loopback mechanism," *IEEE Journal on Selected Areas in Communications*, 2023.

[33] 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.