Premium Quality or Guaranteed Fluidity? Client-transparent DASH-aware Bandwidth Allocation at the Radio Access Network

Stefania Colonnese, *Senior Member, IEEE,* Francesco Conti, Gaetano Scarano, Izhak Rubin, *Life Fellow, IEEE,* Francesca Cuomo, *Senior Member, IEEE*

***Abstract*—In this paper we propose a novel client-transparent Dynamic Adaptive Streaming over HTTP (DASH) -aware bandwidth allocation strategy. The approach, while being application layer transparent, guarantees fluidity to all users but it provides priority-based services to premium users, identified by their Willingness-To-Pay (WTP) profiles. Since different service qualities can be accommodated using WTP, the approach can be extended to XR services, immersive videos, live uplink streaming. To achieve this goal, the allocation problem is firstly formulated as a classical Game Theory problem whose closed form solution is clearly understood in the literature. In a nutshell, the WTP-based, Game Theoretically Optimal Bandwidth Allocation (WTP-GTOBA), firstly satisfies the minimum bandwidth needs of each user and then fairly distributes the residual bandwidth. WTP-GTOBA can be approximately implemented by a greedy, application layer transparent algorithm to be implemented at a DASH-aware network element managing different neighbouring radio access stations. Thereby, the proposed method is suitable for integration of WTP and resource management among multiple service providers and heterogeneous client groups. Numerical simulations carried out under a realistic scenario show that the proposed approach outperforms state-of-the-art application-layer transparent competitors, providing premium quality and/or guaranteed fluidity to different users based on their WTP.**

***Index Terms*—Bandwidth allocation, quality of experience, SAND/DASH, video streaming.**

I. Introduction

VIDEO streaming, already deemed the killer application in next generation mobile networks [1], experienced an unprecedented boost during the 2020 pandemic. In presence of continuously increasing mobile streaming traffic load, novel solutions are needed at the access and the core network [2]–[4] in order to guarantee the desired user's Quality of Experience (QoE).

At the wireless access network, allocating resources for mobile video streaming is challenging because of the large video traffic fluctuations due to the heavy tailed distribution of video packets' sizes [5], [6]. According to the dynamic adaptive streaming based on HTTP (DASH) paradigm, bandwidth limitations, delays, and errors during the transmission result at the application layer into the depletion of the user's video buffer [7] and to possible stall events that may affect the fluidity of the video playout and reduce the users QoE. To cope with the communication channel limits, in DASH the video bitrate is often adapted by application-layer rate control [8], based on partial knowledge of the video traffic [9], or of the channel state [10], [11]. Thereby, several resource allocation methods pursue a cross-layer approach to QoE optimization by video adaptive bitrate (ABR) selection, based on throughput [12], [13], channel [14], buffer [15] or network prediction [16]. The recent ETSI technical specification [17] on server- and network- assisted DASH (SAND) introduces network elements with partial knowledge of the ongoing streaming session and of the client's QoE metrics [18]. Leveraging SAND, multiuser optimization can be addressed [19], [20]. Besides, ABR can be carried out at the edge network, by multi-coordinators and multi-server frameworks [21], [22], on top of video transcoders [23] or proportional fair schedulers supporting premium users [24]. However, the feasibility of these cross-layer solutions requires a strict joint control of radio access and video clients policies. Therefore, these approaches are not suited for frameworks including multiple service providers and heterogeneous client groups.

Application-layer transparent resource allocation methods differ from the above approaches because they pursue QoE without interacting with the video client ABR strategy. This difference is illustrated in Fig.1. Application-layer transparent schemes include scheduling strategies [25], as well as frequency reuse policies [26]. The methods leverage information regarding the ongoing streaming sessions conducted at different users [27] or the video client buffer status for DASH streaming [28]. They achieve relevant QoE goals, such as minimal average buffer depletion for users sharing a limited bandwidth [29]. Still, the impact of DASH aware radio resources managers has not been addressed in the framework of application layer transparent solutions. Furthermore, the importance of QoE differentiation and users' prioritization at wireless access is expected to grow on next generation networks, where mixed reality streaming, live uplink streaming, or 360 degrees streaming may simultaneously take place at different users' clients. In this framework, we address prioritized, application layer transparent, resource allocation to seamlessly

Manuscript received August 31, 2021; revised December 1, 2021; approved for publication by Patrick Seeling, Division II Editor, December 26, 2021.

S. Colonnese, F. Conti, G. Scarano, and F. Cuomo are with Information Engineering, Electronics and Telecom., Sapienza - University of Rome, via Eudossiana, 18 - 00184 - Roma, Italy, email: {stefania.colonnese,franc.conti,gaetano.scarano,francesca.cuomo}@uniroma1.it.

I. Rubin is with Electrical and Computer Engineering Department, University of California at Los Angeles (UCLA), Los Angeles, US, email: rubin@ee.ucla.edu

S. Colonnese is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2021.000046

Creative Commons Attribution-NonCommercial (CC BY-NC).
This is an Open Access article distributed under the terms of Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided that the original work is properly cited.

1229-2370/21/$10.00 © 2022 KICS

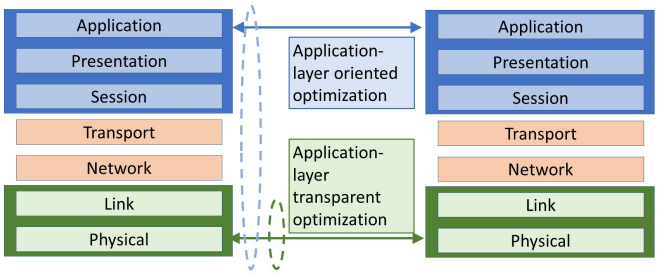


Fig. 1. DASH-aware resource allocation strategies

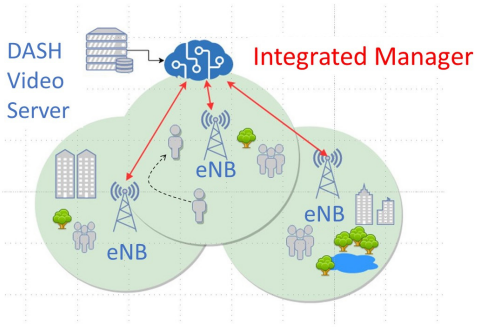


Fig. 2. System Model

provide efficient bandwidth resource management and cope with heterogeneous users. This paper proposes a DASH-aware bandwidth allocation strategy at the wireless access network. The approach jointly addresses the needs of premium and basic users in an optimized way. Specifically, the allocation supports video services with prioritization of a subset of users, while guaranteeing video fluidity to all users. Prioritization is achieved by characterizing each user by a willingness-to-pay (WTP) profile. This prioritization may be related to the user service profile at the radio access network controller, e.g. it can represent the revenue expected by the network operator for the bandwidth allocated to the user. Furthermore, the WTP can be used also to identify institutional users accessing specific services, e.g. immersive video streaming services for law enforcement or crisis management purposes.

The main paper contributions are as follows.

- We recast optimal bandwidth allocation as a classical game theory problem, yielding the WTP-based game theoretically optimal bandwidth allocation (WTP-GTOBA). In a nutshell, the optimal solution is found by i) assigning to each user the minimum bandwidth needed not to stall and ii) fairly sharing the excess bandwidth among the others.
- We show that, under suitable scenarios, WTP-GTOBA is well approximated by a two-stage greedy algorithm (WTP-Greedy), viable for implementation at a DASH aware network element (DANE) and managing different cooperating radio cells.
- WTP-GTOBA, and its greedy approximation WTP-Greedy, support prioritization for a set of users while maintaining fluidity for all the users. This paradigm paves the way to seamless integration of heterogeneous users, i.e. users requiring conventional streaming services with premium users requiring the new services available on next generation networks, where mixed reality, live uplink, or 360 degrees streaming may take place.
- The independence on the adopted ABR policy and the integration of WTP in resource management enables support of multiple service providers and heterogeneous client groups.

The structure of the paper is as follows. In Section II we introduce the system model. Section III presents the WTP-GTOBA and Section IV its greedy version. Numerical performance analysis is reported in Section V and Section VI concludes the paper.

TABLE I
MODEL NOTATION

Symbol	Definition
$\tau[s]$	Chunk duration
$\lambda_k^{(n)}[bits]$	k-th chunk size (n-th user)
K	Chunks' no.
M	No of BSs
$\mathcal{B}^{(m)}$	m-th BS BW [Hz]
N	No of users
$\omega^{(n)}$	n-th user WTP
$\eta_k^{(n)}$ $[(bit/s)/Hz]$	Spectral efficiency (n-th user, k-th time-slot)

II. SYSTEM MODEL

Mobile streaming services typically reflect the MPEG-DASH paradigm, where the client requests, using HTTP, video packets (chunks) from video bitstream pre-encoded at different bitrates and available at the server. The video bitrate selected by the client ABR policy typically depends on the current throughput or client buffer status. We consider N mobile streaming clients, moving within an area covered by a cluster of M radio Base Stations (BSs) [30]. An Integrated Manager, acting as a DANE [31], is used to control and manage the radio resources of the BS cluster (see Fig. 2). Clients are simultaneously streaming video chunks, each of duration τ; for the sake of simplicity and without loss of generality we consider slotted DASH streaming sessions. The k-th chunk $(k = 1, \cdots, K)$ is available at the video servers at video-rates $R_i(Mbit/s), i = 1, \cdots, Q$. The n-th client requires a video chunk at a quality level $q_n^k \in \{1, \cdots, Q\}$, leading to a chunk of length $\lambda_n^k(bits)$. For a given chunk, the link between the user and the serving BS is characterized by a spectral efficiency η $((bit/s)/Hz)$; in order to deliver the k-th video chunk, each user requires a communication channel bandwidth that is equal to: $\lambda_n^k/(\tau \cdot \eta_n^k), \quad n = 1, \cdots, N.$

Each user is associated with a WTP factor ω_n $(\$/Hz)$, modeling the priority of the user by the pay that the client is willing to make for the bandwidth allocated during the service session. A bandwidth B_n^k [Hz] is allocated to the n-th user for the k-th chunk, based on the specific wireless technology adopted in the system (e.g., OFDM in LTE). Table I summarizes the notations.

III. WTP-GTOBA: WTP DEPENDENT GAME THEORETIC OPTIMAL BANDWIDTH ALLOCATION

In the following, we introduce our novel bandwidth allocation as the solution to a game-theoretic problem.

We consider the N clients, served by a cluster of BSs $\mathcal{M}$, as players of a game. For the k-th chunk, the strategy of the n-th client is expressed in terms of its allocated bandwidth B_n^k. Each client specifies a minimum required video quality level, which relates to its WTP level.

We design a client utility function based on the following system relevant parameters:

Buffer status The buffer status $\rho_{b,n}^k \in [0,1)$ is a decreasing function of the buffer occupancy b_k^n:

$$\rho_{b,n}^k = b_{stall}/(b_k^n + 1),$$

where b_{stall} is a buffer occupancy threshold. Large values of $\rho_{b,n}^k \in [0,1)$ indicate that the buffer occupancy is low, and are associated with high risk of stalls.

Upgrade request The pursuit of higher visual quality by means of an upgrade request is accounted for by the binary parameter $\delta_{s,n}^k$, which is equal to 1 if the nth user is requiring a bitrate increase at kth chunk, and 0 otherwise.

Willingness-to-pay the WTP profile of the n-th user in the k-chunk is represented by the constant ω_n.

To this aim, we introduce the utility $u_n^k = \mathcal{U}_n(B_n^k)$ for user n at chunk k. The utility achieved depends on the bandwidth allocated B and the user's QoE parameters ρ_b, δ_s, ω. We herein consider the function $u(B) = \mathcal{U}(B; \rho_b, \omega, \delta_s)$ defined as follows:

$$u(B) = \mathcal{U}(B; \rho_b, \omega, \delta_s) = \frac{B\eta}{\alpha_\rho \rho_b + \alpha_w \omega + \alpha_s \delta_s} = B \cdot \frac{1}{\gamma}, \quad (1)$$

where α_p, α_w and α_s are weighting coefficients and γ compactly summarizes the proportionality factor between the allocated bandwidth and the QoE-related utility function. The definition of the utility in (1) depends on the three relevant QoE factors, namely visual quality, timeliness and prioritization, and it is exploited to derive the solution.

We assume that each user requires a minimum quality level $u_n^{\min}$ that corresponds to the following minimum allocated level:

$$B_n^{\min} = \mathcal{U}^{-1}(u_n^{\min}; \rho_b, \omega, \delta_s) = \gamma_n \cdot u_n^{\min}, \quad (2)$$

where γ_n is the proportionality factor between the minimum n-th user's required utility $u_n^{\min}$, related to the user's QoE parameters, and the minimum allocated bandwidth $B_n^{\min}$. We formulate the following maximization problem of the reward $\mathcal{R}(B_0, \cdots B_{N-1}) = \max \prod_{n=0}^{N-1} \left(u_n^k(B_n) - u_n^{(\min)} \right)$:

$$\begin{aligned} &\max_{B_0, \cdots B_{N-1}} \mathcal{R}(B_0, \cdots B_{N-1}) \\ &= \max_{B_0, \cdots B_{N-1}} \prod_{n=0}^{N-1} \left(u_n^k(B_n) - u_n^{(\min)} \right) \\ &s.t. \sum_n B_n \le B_T, \quad B_n \ge B_n^{\min}, \end{aligned} \quad (3)$$

where B_T denotes the total bandwidth available at the IM level (i.e. over the BSs' cluster) and the utility $u(B)$ is proportional to the amount of resources allocated $u(B) \propto B$. The Nash bargaining solution (NBS) of (3) is defined as the efficient, linear and symmetric solution that is independent from irrelevant alternatives. For the problem in (3), the NBS is known in closed form.

The above formulated problem is a particular case of the following optimization problem: given a budget Ξ for N players, characterized by an individual utility $u_i^{(\min)}$, and a utility versus budget function $u(\xi)$, select the individual budget $\xi_i, i-0, \cdots, N-1$ so as to maximize the overall utility:

$$\begin{aligned} &\Pi_{i=0}^{N-1} \left(u(\xi_i) - \alpha u_i^{(\min)} \right) \\ &\text{subject to } \sum_{i=0}^{N-1} \xi_i \le \Xi, \ \alpha = \max(p \cdot \Xi / \sum_{i=0}^{N-1} u_i, \ 1), \end{aligned} \quad (4)$$

for any $0 < p \le 1$. The solution of this problem is derived in closed form in [32], and extended to the case of a parametric utility function in [33].

For the problem presented in (3), the optimal allocated bandwidth is written as:

$$B_n = B_n^{\min} + \left(B_T - \sum_n B_n^{(\min)} \right) / N,$$

with $B_n^{\min} \propto u_n^{\min}$ and the minimum utility in (3) evaluated as $u_n^{\min} = C \cdot q_n^{\min}$, where $C = C(p)$ is a normalization factor introduced to comply with the constraints in (3) with $\sum_n B_n = p\ B_T, 0 < p \le 1$. Specifically,

$$B_n = \gamma_n^{(p)} q_n + \frac{1}{N} \left(B_T - \sum_n \gamma_n^{(p)} \cdot q_n \right), \quad (5)$$

being

$$\gamma_n^{(p)} = \gamma_n \cdot \min(1, p\ B_T / \sum_i \gamma_n q_n^{\min}),$$

i.e. p scales the bandwidth associated to the minimum quality requirements so as to fit the assigned budget. Thereby, B_n is the sum of two components, namely the bandwidth pB_T needed to support a minimum quality level, and that resulting from a fair division of the $(1-p)B_T$ remaining bandwidth; in Section IV we leverage (5) to derive a two-stage allocation algorithm.

We illustrate the method in a toy case scenario involving $N=2$ users reading the buffer with a fixed playout rate equal to $2Mbit/s$, while the writing rate is $B_n^k \cdot \eta_n^k$, with $\eta_1 = \eta_2 = 1$; no ABR control is performed. We set $B_T = 3.9\ MHz$, $p = 0.5$, $\alpha_r = \alpha_\omega = \alpha_s = 1$, and $q_1^{\min} = q_2^{\min} = 1Mbit/s$. The initial buffer occupancy is the same for the two users, with $b_0^1 = b_0^2 = 6$ chunks. As for WTP, $\omega_1 = 0.8$, $\omega_2 = 0.5$, i.e. the first user is favored.

In Fig. 3(a) we plot the 2D reward function $\mathcal{R}(B_1, B_2) = \left(u_1^k(B_1) - u_1^{(\min)} \right) \cdot \left(u_2^k(B_2) - u_2^{(\min)} \right)$ versus the (B_1, B_2) at the initial chunk $k = 1$. The vertical and horizontal red lines represent the constraints on the minimal allocated bandwidths $B_1 \ge B_1^{\min}$, $B_2 \ge B_2^{\min}$ and the diagonal red line represent the constraint on the overall bandwidth budget $B_1 + B_2 \le B_T$. The maximum reward is obtained by for $B_1 > B_2$ since $\omega_1 > \omega_2$, since all the other conditions are the same for both the users.

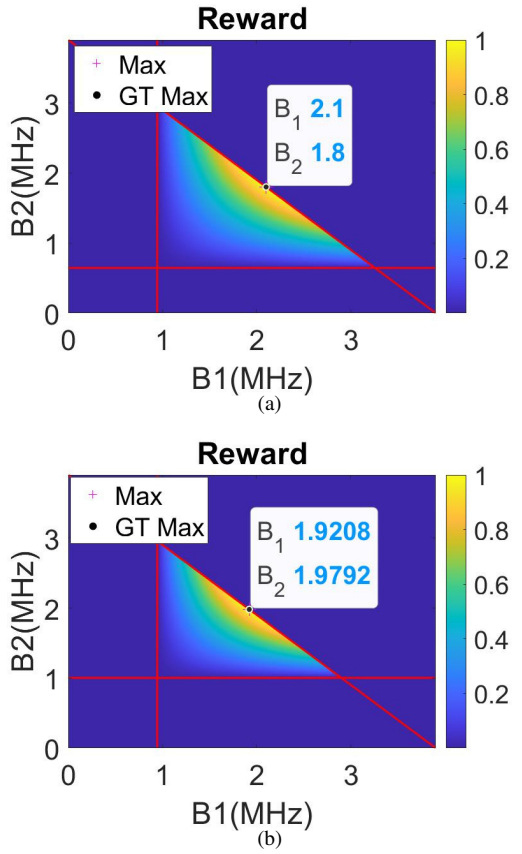


Fig. 3. WTP-GTOBA reward analysis: (a) Reward function at slot $k = 1$ and (b) at slot $k = 50$; red lines represent marginal and total bandwidths constraints.

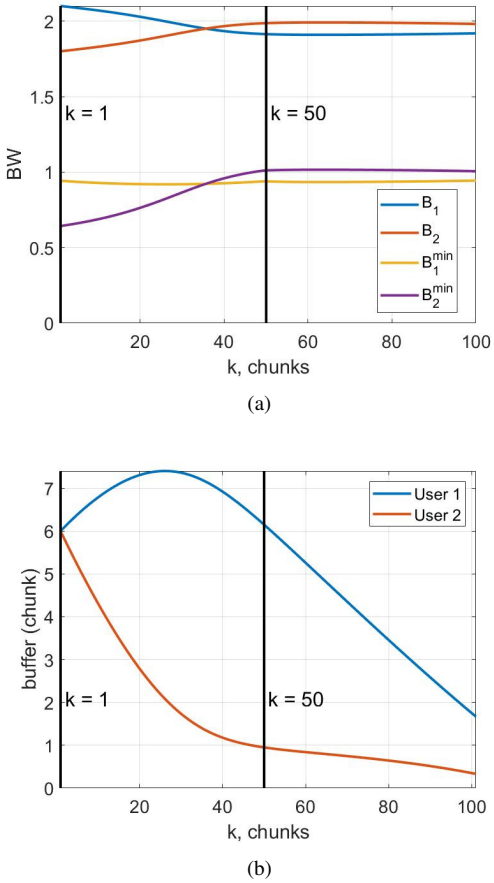


Fig. 4. WTP-GTOBA temporal evolution: (a) Allocated bandwidths (red and blue lines) and minimal bandwidth components (yellow and purple lines); (b) buffer states (red and blue lines).

Then, the system evolves in time. Fig. 4 exhibits the evolution of the allocation in terms of allocated bandwidth (Fig. 4(a)) and buffer status (Fig. 4(b)). Initially, to user 1 is allocated a larger bandwidth because of its higher WTP; this causes an increase of b_k^1 and a decrease of b_k^2. When b_k^2 reduces such that the buffer of user 2 is close to depletion, the maximum of the reward function $\left(u_1^k(B_1) - u_1^{(\min)}\right) \cdot \left(u_2^k(B_2) - u_2^{(\min)}\right)$ shifts. This is exemplified in Fig. 3(b), reporting the reward versus the (B_1, B_2) at chunk $k = 50$. In this condition, the bandwidth allocated to user 2 increases so as to guarantee fluidity. In the final phase of the simulation the two users are granted a balanced allocation.

IV. WTP-Greedy: A Greedy Version of GTOBA

Herein, the WTP-GTOBA bandwidth allocation in (5) is used to define a greedy algorithm to be adopted in a mobile DASH environment. Specifically, we seek for an allocation policy that 1) is partially or totally decentralized, 2) accounts for actual DASH requests, 3) is compatible on a fully granular model for the rate request $R_i, i = 1, \cdots, Q$. Stemming from (5), we present a greedy resource management policy, to which we refer to as WTP-Greedy. Preliminary results about WTP-Greedy appear in [31]. In the following, we show that, in the considered scenario, its two-stage structure provides an accurate approximation of the novel WIP-GTOBA viable for implementation at the DANE.

In (5), we recognize that, apart for possible rescaling due to insufficient bandwidth resources, the optimal solution is split into two terms, corresponding to the bandwidth required to achieve the minimum quality, and to the redistribution of the residual bandwidth, respectively. Thereby, we envisage two allocation stages. Firstly, each user is guaranteed a bandwidth to support a minimum quality level, and secondly the remaining bandwidth is fairly divided among users.

For each video chunk time slot, WTP-Greedy dynamically assigns users to the following QoE-related lists:

1) $\mathcal{L}^{stall}$, users close to a stall event (buffer shortage as measured by comparing $\rho_{b,n}^k$ against a fixed threshold $\rho_{stall} = b_{stall}/(1 + b_{stall})$,
2) $\mathcal{L}^{steady}$, users requiring the same video rate quality as in the previous chunk,
3) $\mathcal{L}^{high}$, users requiring higher video rates.

The WTP-Greedy algorithm is realized by cascading a decentralized and a centralized stage.

At the first stage, clients in $\mathcal{L}^{stall}$ are served with highest priority in a decentralized way by their BS, according to increasing requested bandwidth order. Let us notice that the clients in $\mathcal{L}^{stall}$ are served up to the BS's available bandwidth,

which may or may not suffice to satisfy their requests. Then, if there is any residual bandwidth, the clients in $\mathcal{L}^{steady}$ are served by their BS according to their increasing bandwidth B_n^k and up to the BS's available bandwidth.

The second stage allocates resources in a centralized way to users that were not served at a decentralized level. Any residual BS bandwidth is collected in a shared bandwidth pool B_{pool} available for management at a centralized level. The IM manages the bandwidth pool to serve with high priority the remaining users in $\mathcal{L}^{stall}$ according to their increasing bandwidth B_n^k. Then, the IM allocate the residual bandwidth to the remaining users in $\mathcal{L}^{steady}$, $\mathcal{L}^{high}$, which are served according to their decreasing revenue $B_n^k \cdot \omega_n$. This prioritization guarantees the maximum revenue for a given number of served users. The method can be straightforwardly generalized to the case where the IM manages the shared pool plus an additional bandwidth amount [31].

In the following, we show that the decentralized and centralized stages of the WTP-Greedy algorithm, roughly providing the minimum service level (reduced stall duration) and then redistributing the remaining bandwidth among the other users, according to a WTP based criteria, results in an accurate approximation of the two-components bandwidth allocated WTP-GTOBA in the considered scenario. Therefore WTP-Greedy can act as a proxy of WTP-GTOBA, but it can be feasibly integrated into a DASH environment, and specifically it can be implemented at the IM level. The approach scales with the number of users since it implements the per-chunk optimization by segmenting and sorting the users lists. Different from several state-of-the-art approaches, WTP-greedy is transparent to the application layer ABR policy.

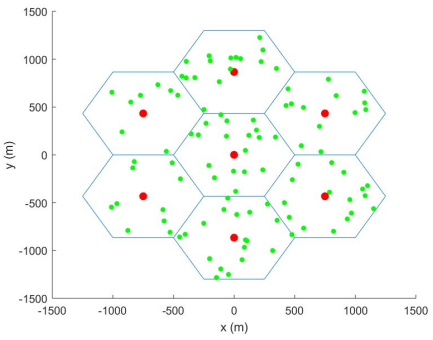
A remark on the additional signaling overhead of the WTP-greedy algorithm is in order. The algorithm requires the back and forth signaling of the unserved requests between the decentralized and centralized levels, as well the signaling of the residual resources from the decentralized to the centralized level. It is worth noting that the signaling is carried out at the application layer time scale, i.e. every few seconds, and it is therefore much looser than typical link layer signaling. On the other hand, an additional computational cost is required at the IM for optimally managing the resources; still, this goes in the way of incrementing the intelligence in network resource managing in next generation cellular networks [34].

V. Simulations and Results

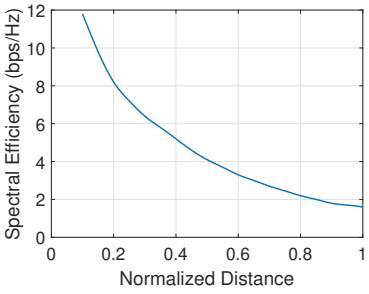
In this section, we analyse the performance of the optimal WTP-GTOBA (5) and of the WTP-Greedy algorithm. Firstly, we show that WTP-GTOBA is approximated by WTP-Greedy in the considered scenario. Then, we assess the performance of this latter in comparison with state-of-the-art in radio resources management for mobile streaming.

We show that the WTP-Greedy algorithm approximates the optimal WTP-GTOBA in (5) in the mobile video streaming scenario simulated as follows.

For concreteness sake, we consider the scenario in Fig. 5, encompassing $M = 7$ hexagonal BS cells. This can be seen as a schematic representation of the resource sharing



(a)



(b)

Fig. 5. Simulation Scenario: (a) Network scenario and (b) spectral efficiency.

architecture Fig. 2, where the IM manages the radio resources of a BS cluster. Besides, it can be also deemed as an abstract representation of future architecture involving multipoint coordination [35], [36] which are currently under investigation in 5G [37] and 6G networks [38]. The adaptation of the scenario parameter setting to future generation networks is left for further study.

The overall bandwidth available at each BS is equal to $\mathcal{B}_m = 10\ MHz$. During the session $N = 75$ users move in a direction and at a speed randomly selected in $[0, 2\pi]$ and $[0, 5\,m/s]$, respectively (bounce conditions are further inserted to prevent users from leaving the area of interest). The users are randomly assigned 5 WTP equally likely profiles, identified by the values $\omega = [0.2, 0.4, 0.6, 0.8, 1]\ \$/Hz$. We consider a slotted DASH session.

Synthetic 360-degrees video traffic traces are generated according to the model in [39]. The traces refer to $Q = 5$ quality levels at rates $[0.55, 1.06, 1.95, 4.01, 8.46]$ (Mbps); the chunk sizes $\lambda_n^k(\text{bits})$ are generated using a Gamma distribution, whose parameters are summarized in Table II. Each user's client implements a buffer-based ABR policy, out of the control of the DANE. ABR policies as BOLA [40] require solving an optimization step at each chunk. For the sake of lowering the computational complexity, we implemented a computationally lighter hybrid throughput/buffer based ABR[1]. If a stall occurs, the client buffer is refilled in a fixed amount

[1]After been served for 3 time slots, the user requires a higher quality; after having not been served in 2 out of 4 time slots, the user requires a lower quality; if the user's buffer is lower than a critical threshold b_{stall} the lowest available rate is selected.

TABLE II
VIDEO PARAMETERS

	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
R_i (Mbps)	0.55	1.06	1.95	4.01	8.46
α	12.13	12.21	11.71	12.29	15.75
β	73850	140059	269005	526355	867408

of time; therefore, the number of stalls is proportional to the overall stall duration. The bandwidth request is computed as: $B_n^k = \lambda_n^k/(\tau \cdot \eta_n^k), \quad n = 1, \cdots, N.$

As for the mobile network model, we built a channel model based on [41]. The spectral efficiency is set according to the distance between user and BS as in Fig. 5(b). Since we assess the performance of the optimal resource allocation at the radio access network, during the simulation we assume an ideal, constant delay, data transfer in the core network. For each chunk, the allocated bandwidth is computed according to the WTP-Greedy allocation policy. Furthermore, the WTP-GTOBA optimization is carried out on a chunk-by-chunk basis, and the WTP-GTOBA bandwidth is computed as in (5), The parameters $\alpha_r, \alpha_\omega, \alpha_s$ represent the weights of the buffer status ρ_b, the willingness-to-pay ω, and the upgrade request δ_s respectively. They are set as $\alpha_r = 1$, $\alpha_\omega = 4$, $\alpha_s = 1$. Let us observe that the resource allocation could be molded by different objectives, like preventing stalls by increasing the parameter α_r on one hand, or discouraging up-switching by increasing the parameter α_δ on the other hand. For the sake of compactness, we restrict ourselves to the case $\alpha_\omega > \alpha_s = \alpha_r$, so as to highlight the effect of the willingness-to-pay parameter ω on the overall allocation. In the following we illustrate the comparison of the WTP-GTOBA optimal allocated bandwidth with that allocated by WTP-greedy. The results have been obtained by setting $p = 0.8$, i.e. allocating up to 80% of the available bandwidth for minimum quality provisioning and the remaining 20% for further redistribution. This implies that 80% of the bandwidth is distributed by accounting for the current user QoE parameters whereas the remaining 20% is distributed in a flat way. It is worth remarking that the proposed allocation strategy boils down to uniform resource allocation for $p = 0$, and in general it can be adapted to different allocation criteria by suitable parameter settings.

Fig. 6 shows the boxplot, over 30 runs, of the bandwidths averaged over $K = 100$ chunks. We compare the WTP-GTOBA and the WTP-greedy average bandwidth; for both the methods, the boxplot shows that the per run deviations from the average are small. We recognize that WTP-greedy tightly approaches the WTP-GTOBA allocated bandwidth. For different WTPs the approximation is from below or above, also given that the WTP-Greedy allocated bandwidths for each chunk directly reflect the quantization of the users bitrate requests while the WTP-GTOBA allocated bandwidths vary in a continuous range. The results in Figure 6 clearly show that the resource distribution by WTP-Greedy well approximates the optimal WTP-GTOBA allocation in the considered scenario. The reason why this occurs is to be found in the analogy between the two approaches, which split the allocation into two conceptual stages, i.e. the satisfaction of minimal QoE

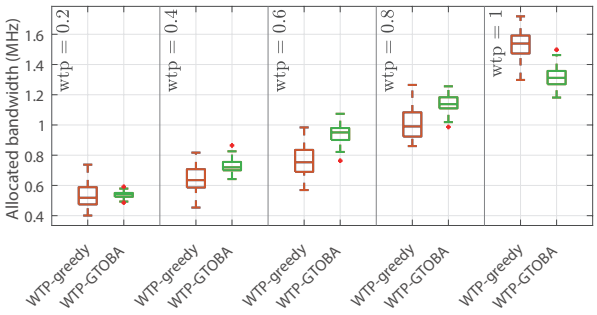


Fig. 6. WTP-Greedy vs WTP-GTOBA.

constraints, and the fair allocation of the remaining resources.

In the following, we illustrate by further simulations that this guarantees prioritization to some users and fluidity to all the users. In Fig. 7, we display the main QoE metrics averaged over $nRuns = 30$ runs, for each group of users, namely:
Video quality, measured as the average playout rate of the received chunks

$$\mathcal{Q} = \frac{1}{KN} \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} q_k^{(n)};$$

Video fluidity, measured as

$$\mathcal{F} = K - \overline{n_{stalls}},$$

where $\overline{n_{stalls}}$ is number of equal-duration stalls experienced by each user during the session;
Video smoothness, measured as the percentage of chunks over which the quality does not decrease:
Fairness, computed as the Jain's fairness index on video-quality of users belonging to the same WTP profile, i.e. characterized by the same WTP value:

$$\mathcal{J}_{\overline{\omega}} = \frac{\left(\frac{1}{KN} \sum\limits_{n \text{ s.t. } \omega_n = \overline{\omega}} \sum\limits_{k=0}^{K-1} q_k^{(n)} \right)^2}{\frac{1}{N} \sum\limits_{n \text{ s.t. } \omega_n = \overline{\omega}} \left(\frac{1}{K} \sum\limits_{k=0}^{K-1} q_k^{(n)} \right)^2}.$$

It can be seen that WTP-Greedy achieves a good users differentiation in QoE guaranteeing a high fluidity of the service also to users with lower willingness-to-pay.

We now compare WTP-Greedy with proportional fair (PF) [25], proportional fair quality-aware (PF-QAW) [28] and minimum average delay (MAD) [29] resource allocation algorithms. Let us first consider the case of uniform WTP profiles.

The PF approach is a classic resource allocation method used in LTE BSs, where the down-link radio resources are organized in resource blocks (RB), which is a time-frequency cell of 180 kHz bandwidth for a 1 ms transmission time interval (TTI). In PF each RB is assigned to the user maximizing a factor proportional to the expected data rate and inversely proportional to the average of the past throughput. The QoE scheduler in [28] modifies the factor using estimates of the user's buffer in DASH streaming, trying to serve with higher priority the users close to the stall. By contrast,

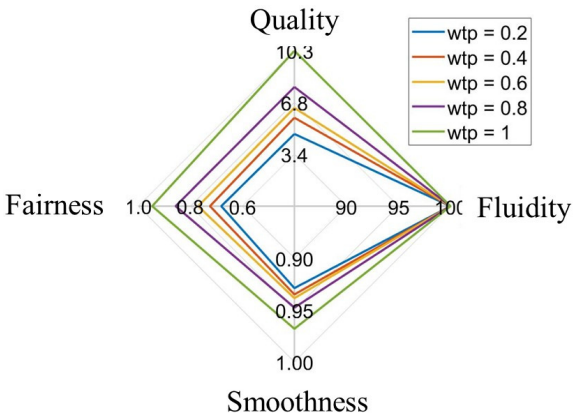


Fig. 7. Performance for different WTPs.

TABLE III
PERFORMANCE COMPARISON (DIFFERENT N)

	$\overline{q}$ $(Mbps)$				$\overline{n_{stalls}}$			
N	WTP-Greedy	MAD	PF	PF-QAW	WTP-Greedy	MAD	PF	PF-QAW
75	**7.31**	7.00	7.23	7.21	**0**	2.75	3.07	2.95
100	**5.83**	5.38	5.64	5.57	**0**	3.16	3.85	3.66
125	**4.91**	4.44	4.68	4.61	**0**	3.38	4.39	4.12
150	**4.05**	3.66	3.88	3.80	**0**	3.70	5.10	4.83

TABLE IV
PERFORMANCE COMPARISON (DIFFERENT WTP ω)

	$\overline{q}$ $(Mbps)$				$\overline{n_{stalls}}$			
ω	WTP-Greedy	MAD	PF	PF-QAW	WTP-Greedy	MAD	PF	PF-QAW
0.2	5.26	5.36	**5.63**	5.56	**0**	3.16	3.83	3.61
1	**6.61**	5.49	5.76	5.69	**0**	3.16	3.85	3.65

the paper [29] proposes a cross-layer chunk-wise allocation based on a reverse water-filling principle, aiming to minimize the average chunk delays. We have integrated these policies into the simulator and analyzed the performance for different values of N and K. Let us observe that WTP-Greedy carries out the transmission of a chunk or not, acting as an admission control procedure. On the contrary, competing policies serve all users, even if the bandwidth may not be sufficient to transmit a whole chunk.

The performance indices in terms of video quality and number of stalls, averaged over 30 runs, are summarized in Table III. The results should be read as follows: the PF method is DASH unaware; PF-QoE method, accounting for buffer status, avoids some stalls at the expense of resource efficiency. Similarly, the MAD method [29] avoids a larger number of stalls by sacrificing more video quality. The WTP-Greedy approach is able to avoid stalls caused by radio access network bottleneck events, gaining in terms of resource efficiency. This is due to the chunk-wise admission control procedure that leads to serving users close to the b_{stall} with the highest priority.

Finally, we repeat the experiment equally dividing $N = 100$ users in two groups with 2 different WTPs, $\omega_1 = 0.2$ \$/Hz and $\omega_2 = 1$ \$/Hz. The results exhibited in Table IV confirm the effectiveness of the proposed approach to include quality prioritization between the two classes and quantify the amount of video quality difference in comparison with methods that do not prioritize users. These results have been achieved without introducing video stalls (more precisely, the stalls due to radio access) affecting low WTP users. For comparison sake, we have reported the performance of state of the art alternatives, noting that these schemes do not support QoE differentiation.

VI. CONCLUSIONS

In this paper, we propose an application-layer transparent, priority based, bandwidth allocation method for mobile video streaming. We formulate the problem targeting both user quality of experience requirements and revenue-related parameters based on willingness-to-pay (WTP). Then, we derive a game theoretically optimal bandwidth allocation (WTP-GTOBA) algorithm, and we show that it admits a greedy version (WTP-greedy) that can be implemented at a DASH aware network element managing radio resources. Simulations show that WTP-GTOBA and its approximation by the WTP-Greedy algorithm outperform state-of-the-art alternative schemes. The proposed approach integrates users prioritization thus paving the way for accommodating services at different qualities, such as XR services, immersive videos, live uplink streaming. Furthermore, being application layer transparent, it is suited to multiple service providers and heterogeneous client groups.

REFERENCES

[1] H. Jung, "Cisco visual networking index: Global mobile data traffic forecast update 2017–2022 white paper," *Tech. Rep, Cisco Systems Inc.*, 2019.
[2] A. Feldmann *et al.*, "The lockdown effect," in *Proc. ACM IMC*, 2020.
[3] A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, and J. Khangosstar, "A characterization of the covid-19 pandemic impact on a mobile network operator traffic," in *Proc. ACM IMC*, 2020.
[4] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-learning during covid-19 pandemic," *Comput. Netw.*, vol. 176, p. 107290, Jul. 2020.
[5] I. Rubin, S. Colonnese, F. Cuomo, F. Calanca, and T. Melodia, "Mobile http-based streaming using flexible LTE base station control," in *Proc. IEEE WoWMoM*, 2015.
[6] L. Rossi, J. Chakareski, P. Frossard, and S. Colonnese, "A poisson hidden markov model for multiview video traffic," *IEEE/ACM Trans Netw.*, vol. 23, no. 2, pp. 547–558, Feb. 2014.
[7] S. Colonnese, S. Russo, F. Cuomo, T. Melodia, and I. Rubin, "Timely delivery versus bandwidth allocation for dash-based video streaming over LTE," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 586–589, Jan. 2016.
[8] J. Kua, G. Armitage, and P. Branch, "A survey of rate adaptation techniques for dynamic adaptive streaming over http," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, no. 3, pp. 1842–1866, 2017.
[9] S. Colonnese, F. Cuomo, T. Melodia, and R. Guida, "Cloud-assisted buffer management for http-based mobilevideo streaming," in *Proc. ACM PE-WASUN*, 2013.
[10] A. Dvir, N. Harel, R. Dubin, R. Barkan, R. Shalala, and O. Hadar, "Misal-a minimal quality representation switch logic for adaptive streaming," *Multimedia Tools Applications*, vol. 78, no. 18, no. 18, pp. 26483–26508, 2019.
[11] S. Colonnese, F. Cuomo, K. Miller, V. Sapio, and A. Wolisz, "Affordable delay based quality selection for http adaptive video streaming," in *Proc. IEEE LANMAN*, 2017.
[12] K. Spiteri, R. Urgaonkar, and R. Sitaraman, "Bola: Near-optimal bitrate adaptation for online videos," in *Proc. IEEE INFOCOM*, 2016.

[13] T. Mangla, N. Theera-Ampornpunt, M. Ammar, E. Zegura, and S. Bagchi, "Video through a crystal ball: Effect of bandwidth prediction quality on adaptive streaming in mobile environments," in *Proc. SIGMM MoVid*, 2016.
[14] Z. Lu and G. De Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE INFOCOM*, 2013.
[15] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. ACM SIGCOMM*, 2014.
[16] F. Bronzino, D. Stojadinovic, C. Westphal, and D. Raychaudhuri, "Exploiting network awareness to enhance dash over wireless," in *Proc. IEEE CCNC*, 2016.
[17] "Study on server and network-assisted dynamic adaptive streaming over http (dash) (sand) for 3GPP multimedia services," *3GPP TR 26.957 Release 16*, Jul. 2020.
[18] H. Bermudez, J. Martinez-Caro, R. Sanchez-Iborra, J. Arciniegas, and M. Cano, "Live video-streaming evaluation using the itu-t p. 1203 qoe model in LTE networks," *Comput. Netw.*, vol. 165, p. 106967, Dec. 2019.
[19] V. Nathan, V. Sivaraman, R. Addanki, M. Khani, P. Goyal, and M. Alizadeh, "End-to-end transport for video qoe fairness," in *Proc. ACM SIGCOMM*, 2019.
[20] I. Triki, R. El-Azouzi, and M. Haddad, "Newcast: Joint resource management and qoe-driven optimization for mobile video streaming," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 2, pp. 1054–1067, Jun. 2020.
[21] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski, "Joint optimization of qoe and fairness through network assisted adaptive mobile video streaming," in *Proc. IEEE WiMob*, 2017.
[22] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski, "Edge computing assisted adaptive mobile video streaming," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 787–800, Apr. 2018.
[23] Y. Guo, F. Yu, J. An, K. Yang, C. Yu, and V. Leung, "Adaptive bitrate streaming in wireless networks with transcoding at network edge using deep reinforcement learning," *IEEE Trans. Vehicular Technol.*, vol. 69, no. 4, pp. 3879–3892, Apr. 2020.
[24] X. Xu, R. Govindan, A. Mahimkar, N. Shankaranarayanan, J. Wang, and M. Yu, "Enabling premium service for streaming video in cellular networks," in *Proc. IEEE IFIP*, 2020.
[25] I. Sousa, M. Queluz, and A. Rodrigues, "A survey on qoe-oriented wireless resources scheduling," *J. Netw. Comput. Applications*, vol. 158, p. 102594, May 2020.
[26] H.-B. Chang, I. Rubin, S. Colonnese, F. Cuomo, and O. Hadar, "Joint adaptive rate and scheduling for unicasting video streams in cellular wireless networks," *IEEE Trans. Vehicular Technol.*, vol. 66, no. 9, pp. 8398–8412, Sep. 2017.
[27] D. Perdana, A. Sanyoto, and Y. Bisono, "Performance evaluation and comparison of scheduling algorithms on 5G networks using network simulator," *Int. J. Comput. Commun. Control*, vol. 14, no. 4, pp. 530–539, Aug. 2019.
[28] J. Navarro-Ortiz, P. Ameigeiras, J. M. Lopez-Soler, J. Lorca-Hernando, Q. Perez-Tarrero, and R. Garcia-Perez, "A qoe-aware scheduler for http progressive video in ofdma systems," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 677–680, Apr. 2013.
[29] 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 Commun. Lett.*, vol. 21, no. 2, pp. 386–389, Feb. 2016.
[30] A. Zubow, A. Rostami, and S. Bayhan, "On practical cooperative multi point transmission for 5G networks," *Comput. Netw.*, vol. 171, p. 107105, Apr. 2020.
[31] F. Conti, S. Colonnese, F. Cuomo, L. Chiaraviglio, and I. Rubin, "Quality of experience meets operators revenue: Dash aware management for mobile streaming," in *Proc. IEEE EUVIP*, 2019.
[32] I. Ahmad and J. Luo, "On using game theory to optimize the rate control in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 209–219, Feb. 2006.
[33] S. Colonnese, G. Panci, S. Rinauro, and G. Scarano, "Optimal video coding for bit-rate switching applications: a game-theoretic approach," in *Proc. IEEE WoWMoM*, 2007.
[34] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent o-ran for beyond 5G and 6G wireless networks," *arXiv preprint arXiv:2005.08374*, 2020.
[35] F. Z. Morais, C. A. da Costa, A. M. Alberti, C. B. Both, and R. da Rosa Righi, "When SDN meets C-RAN: A survey exploring multipoint coordination, interference, and performance," *J. Netw. Comput. Applications*, vol. 162, p. 102655, Jul. 2020.
[36] M. S. J. Solaija, H. Salman, A. B. Kihero, M. İ. Sağlam, and H. Arslan, "Generalized coordinated multipoint framework for 5G and beyond," *IEEE Access*, vol. 9, pp. 72499–72515, May 2021.
[37] S. Zaidi, O. B. Smida, S. Affes, U. Vilaipornsawai, L. Zhang, and P. Zhu, "User-centric base-station wireless access virtualization for future 5G networks," *IEEE Trans. Commun.*, vol. 67, pp. 5190–5202, Jul. 2019.
[38] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.
[39] S. Colonnese, F. Cuomo, L. Ferranti, and T. Melodia, "Efficient video streaming of 360 cameras in unmanned aerial vehicles: an analysis of real video sources," in *Proc. IEEE EUVIP*, 2018.
[40] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "Bola: Near-optimal bitrate adaptation for online videos," *IEEE/ACM Trans. Netw.*, vol. 28, no. 4, pp. 1698–1711, Aug. 2020.
[41] A. Hajisami and D. Pompili, "Dynamic joint processing: Achieving high spectral efficiency in uplink 5G cellular networks," *Comput. Netw.*, vol. 126, pp. 44–56, Oct. 2017.

Stefania Colonnese is currently Associate Professor at the Department of Information Engineering, Electronics and Telecommunications (DIET) of Sapienza University of Rome, Italy. She received her Ph.D. in Electronics Engineering from the University of Roma Tre. She has participated in the MPEG-4 standardization activity within the ISO MPEG-4 Core Experiment on Automatic Video Segmentation. She is co-author of more than a hundred journal and conference papers, two book chapters and several ISO MPEG-4 Contributing Documents. Her research interests range from statistical signal processing, image deconvolution and restoration, biomedical signal processing, to video encoding, processing and networking. She served in several conferences as Technical Program Co-chair (IEEE/Eurasip EUVIP 2014), Publicity Chair, Technical Program Committee member. She has served as Associate Editor of the Hindawi International Journal of Digital Multimedia Broadcasting (IJDMB), devoted to the topics of Multimedia Broadcasting, Standardization, and Quality of Experience. She is an IEEE Senior Member.

Francesco Conti was born in Pontecorvo, Italy. He received the Bachelor and Master Degree in Communication Engineering from the University of Roma La Sapienza in 2017 and 2020 respectively. He held a Research Fellowship in signal and image processing at the Department of Computing Sciences, Tampere University, Finland. He held a research fellowship at the Department of Information Engineering, Electronics and Telecommunications (DIET) of Sapienza University of Rome, Italy, in 2020. He is currently employed at MBDA Italy as Image Processing Engineer in the "Guidance, Control and Navigation Departement".

Gaetano Scarano was born in Campobasso, Italy. He received the "Laurea" degree in Electronic Engineering (summa cum laude) from Università di Roma "La Sapienza," Rome, Italy, in 1982. In 1982, he joined the Istituto di Acustica, Consiglio Nazionale delle Ricerche, Roma, Italy, as Associate Researcher. Since 1988, he has been teaching digital signal processing at the University of Perugia, Perugia, Italy, where in 1991 he became Associate Professor of Signal Theory. In 1992, he joined the Dipartimento di Scienza e Tecnica dell'Informazione e della Comunicazione, now Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni, Università di Roma "La Sapienza," first as Associate Professor of Image Processing, then as Professor of Signal Theory. His research interests lie in the area of signal and image processing, communications, estimation and detection theory, and include channel equalization and estimation, image restoration, and texture synthesis and classification. He served as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.

Izhak Rubin received the B.Sc. and M.Sc. from the Technion - Israel Institute of Technology, Haifa, Israel, in 1964 and 1968, respectively, and the Ph.D. degree from Princeton University, Princeton, NJ, in 1970, all in Electrical Engineering. Since 1970, he has been on the faculty of the UCLA School of Engineering and Applied Science where he is a Distinguished Professor in the Electrical and Computer Engineering Department. Dr. Rubin has had extensive research, publications, consulting, and industrial experience in the design and analysis of commercial and military computer communications and telecommunications systems and networks. Such design and analysis projects include network systems employed by the FAA for air traffic control, terrestrial and satellite based mobile wireless networks, high speed multimedia telecommunications networks, advanced cellular cross-layer operations, mobile backbone ad hoc wireless networks, mechanisms to assure network resiliency and automatic failover operations, networking and traffic management for autonomous highway transportation systems, UAV aided mobile robust ad hoc wireless networks, and adaptive combined source and channel coding and scheduling of video streaming operations over wireless networks. At UCLA, he is leading a research group in the areas of telecommunications and computer communications networks and serves as co-director of the UCLA Public Safety Network Systems Laboratory and director of the Autonomous Intelligent Networked Systems Laboratory. During 1979–1980, he served as Acting Chief Scientist of the Xerox Telecommunications Network. He served as co-chairman of the 1981 IEEE International Symposium on Information Theory; as program chairman of the 1984 NSF-UCLA workshop on Personal Communications; as program chairman for the 1987 IEEE INFOCOM conference; and as program co-chair of the IEEE 1993 workshop on Local and Metropolitan Area networks, as program co-chair of the 2002 first UCLA/ONR Symposium on Autonomous Intelligent Networked Systems (AINS), and has organized many other conferences and workshops. He has served as an editor of the IEEE Transactions on Communications, Wireless Networks journal, Optical Networks magazine, IEEE JSAC issue on MAC techniques, Communications Systems journal, Photonic Networks Communications journal, and has contributed chapters to texts and encyclopedia on telecommunications systems and networks. Dr. Rubin is a Life Fellow of IEEE.

Francesca Cuomo received the Ph.D. in Information and Communications Engineering in 1998 from Sapienza University of Rome. From 2005 to October 2020 she was Associate Professor and from November 2020 she joined "Sapienza" as Full Professor teaching courses in Telecommunication Networks. She has advised numerous master students in computer in computer engineering, and has been the advisor of 13 Ph.D. students in Networking. Her current research interests focus on: Vehicular networks and Sensor networks, Low Power Wide Area Networks and IoT, 5G Networks, Multimedia Networking, Energy saving in the Internet and in the wireless system. Francesca Cuomo has authored over 154 peer-reviewed papers published in prominent international journals and conferences. Her Google Scholar h-index is 30, $>$3850 citations. Relevant scientific international recognitions: 2 Best Paper Awards. She has been in the editorial board of Computer Networks (Elsevier) and now is member of the editorial board of the Ad-Hoc Networks (Elsevier), IEEE Transactions on Mobile Computing, Sensors (MDPI), Frontiers in Communications and Networks Journal. She has been the TPC co-chair of several editions of the ACM PE-WASUN workshop, TPC Co-Chair of ICCCN 2016, TPC Symposium Chair of IEEE WiMob 2017, General Co-Chair of the First Workshop on Sustainable Networking through Machine Learning and Internet of Things (SMILING), in conjunction with IEEE INFOCOM 2019; Workshop Co-Chair of AmI 2019: European Conference on Ambient Intelligence 2019. She is IEEE senior member.