Multi-User Goal-Oriented Communications With Energy-Efficient Edge Resource Management

Francesco Binucci, *Graduate Student Member, IEEE*, Paolo Banelli, *Member, IEEE*, Paolo Di Lorenzo, *Senior Member, IEEE*, and Sergio Barbarossa, *Fellow, IEEE*

***Abstract*—Edge Learning (EL) pushes the computational resources toward the edge of 5G/6G network to assist mobile users requesting delay-sensitive and energy-aware intelligent services. A common challenge in running inference tasks from remote is to extract and transmit only the features that are most significant for the inference task. From this perspective, EL can be effectively coupled with goal-oriented communications, whose aim is to transmit only the information *relevant* to perform the inference task, under prescribed accuracy, delay, and energy constraints. In this work, we consider a multi-user/single server wireless network, where the users can opportunistically decide whether to perform the inference task by themselves or, alternatively, to offload the data to the edge server for remote processing. The data to be transmitted undergoes a goal-oriented compression stage performed using a convolutional encoder, jointly trained with a convolutional decoder running at the edge-server side. Employing Lyapunov optimization, we propose a method to jointly and dynamically optimize the selection of the most suitable encoding/decoding scheme, together with the allocation of computational and transmission resources, across all the users and the edge server. Extensive simulations confirm the effectiveness of the proposed approaches and highlight the trade-offs between energy, latency, and learning accuracy.**

***Index Terms*—Edge learning, goal-oriented communications, Lyapunov stochastic optimization, deep learning.**

I. Introduction

THE ADVENT of the fifth/sixth generation of mobile communications has radically changed the network concept, from a pure communication infrastructure to a key enabler for pervasive services, which are highly based on Artificial Intelligence (AI) and Machine Learning (ML). Typical examples can be found in augmented reality, autonomous driving, massive Internet of Things, and mission critical applications [1]. In these scenarios, the service delay and the reliability constraints are often very restrictive, and this motivates the need to design a holistic system where communication, computation, learning, and control are jointly managed in order to reach reliability, energy efficiency, and sustainability.

The need to process a huge amount of data, in real-time, through proper AI/ML techniques, has driven researchers to design training/inference tasks at the wireless edge, in collective as well as distributed fashions. This has led to the definition of the so called Edge Intelligence (EI) paradigm [2]. In this view, the allocation of system resources in order to reach prescribed target performance in terms of latency, accuracy, and energy consumption has been already considered in [3], [4], [5], [6]. Specifically, EI allows User Equipments (UEs) connected to a mobile network to opportunistically offload their learning tasks to Edge Servers (ESs), which are placed in the network edge, nearby the Radio Access Points (RAPs). This allows the efficient management of system resources, such as transmission rate, bandwidth, and CPU clock rates, according to specific optimization strategies, which are mainly focused on the trade-offs between energy consumption, overall latency, and learning accuracy [6].

Clearly, in a resource optimization perspective, it would be useful to offload to the ESs only the (minimum) amount of information strictly necessary to fulfill the learning task with the desired accuracy, while respecting the performance requirements. This intuitive consideration, jointly with the huge increase of traffic envisaged in future 6G networks [7], motivates the search for a new communication paradigm, alternative to the classical Shannon design. In this view, a valuable candidate is represented by Goal-Oriented Communications (GOC) [8]. More specifically, if the goal of communication is to perform an inference task on the data collected by the UE, rather than requiring the accurate reproduction of all the transmitted bits at the receiver side, the aim of GOC is to transmit only the information that is most relevant to run the inference task at the ES, guaranteeing a prescribed level of decision accuracy and system performance. In this way, it is possible to help the UEs to save transmission resources and avoid unnecessary data rate growth, still respecting

Manuscript received 1 February 2023; revised 6 April 2023; accepted 2 May 2023. Date of publication 11 May 2023; date of current version 22 November 2023. This work was supported in part by the Fondo di Ricerca di Base 2021 Project APE6G funded by the University of Perugia; in part by the PRIN 2017 Project LIQUID EDGE funded by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) under Grant 2017TRRZY7; and in part by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" through Program "RESTART" under Grant PE00000001. The work of Paolo Di Lorenzo and Sergio Barbarossa was supported by the SNS-JU-2022 Project ADROIT6G under Agreement 101095363. The editor coordinating the review of this article was M. Chen. *(Corresponding author: Francesco Binucci.)*

Francesco Binucci is with the Department of Engineering, University of Perugia, 06125 Perugia, Italy, and also with the Department of Engineering, Consorzio Nazionale Interuniversitario per le Telecomunicazioni-University of Perugia, 06125 Perugia, Italy (e-mail: francesco.binucci@studenti.unipg.it).

Paolo Banelli is with the Department of Engineering, University of Perugia, 06125 Perugia, Italy (e-mail: paolo.banelli@unipg.it).

Paolo Di Lorenzo and Sergio Barbarossa are with the Department of Information Engineering, Electronics, and Telecommunications, Sapienza University of Rome, 00184 Rome, Italy (e-mail: paolo.dilorenzo@uniroma1.it; sergio.barbarossa@uniroma1.it).

Digital Object Identifier 10.1109/TGCN.2023.3275199

application constraints, such as service delay and energy consumption.

Related works. Seminal EI frameworks, with a wireless offloading strategy, have been proposed in [6], [9], which save transmission resources by simply allocating, in a dynamic fashion, the number of (quantization) bits used by UEs to transmit their data to the ES. This compression strategy has also been employed in [10] and [11], where edge classification and ensemble learning are considered, respectively, with reliability guarantees. A more principled data reduction strategy, better matched to the learning task and based on the Information Bottleneck (IB) [12], [13], has been proposed in [14]. However, the IB principle admits a closed form solution for the encoder only if the overall statistics are jointly Gaussian [14], [15], or a solution achievable through an iterative mechanism, if the statistics are discrete. When the sensed data and decision outputs are neither jointly Gaussian, nor discrete with manageable cardinality, it is not easy to derive the IB solution and the source encoding problem can be reformulated using the so called variational IB (VIB), as recently explored in [16] and in [17], where a cooperative (multi-device) inference framework is proposed.

A possibility to further deviate from the classical communication design is offered by Joint Source Channel/Coding (JSCC), which has received increasing attention with the wide spread use of Deep Neural Networks (DNNs). Quite recently, several works have proposed to replace the classical cascade of source and channel encoders with a DNN properly trained with respect to the specific task. For instance, [18] proposed a DNN-based JSCC scheme to achieve higher performance in finite block-length regime for image retrieval applications. Furthermore, if the task of communication is image recognition, it makes sense to design the JSCC architecture directly focusing on the learning task, rather than on the image reconstruction followed by the recognition task, as proposed in [19]. The authors of [20] presented a scheme for image retrieval where the extracted vector features are directly mapped to the channel input symbols, without resorting to any channel coding technique, and the server retrieves the most relevant images directly from the noisy channel output. This approach has been extended in [21], where the extracted features are quantized before being mapped onto the channel symbols. In [22], JSCC is coupled with an OFDM system and operating over a frequency-selective channel, while [23] considers the combination of JSCC with non-linear transform coding [24].

As far as goal-oriented (also known as task-oriented) communications is concerned, several recent works testify the emerging relevance of this topic. For instance, in [25] and [26] GOCs have been exploited to define the *common-language* between a listener and a speaker, employing Reinforcement Learning (RL) and Curriculum Learning (CL), while a transformer-based approach has been proposed to assist image and text transmissions [27]. A noise-aware JSCC for text-transmission is described and assessed in [28], while [29] exploited a hybrid automatic repeat request (HARQ) scheme to improve reliability in sentence semantic transmission. Other examples of image classification for Unmanned Aerial Vehicle (UAV) applications, and a GOC-assisted Visual Question Answering (VQA) task, can be found in [30] and [31], respectively. Furthermore, [19] and [20] motivate the use of GOC schemes for computer vision applications, by showing the accuracy improvements they provide in image-classification and re-identification tasks of humans and cars, respectively. Finally, the impact of goal-oriented communications has also been analyzed in speech recognition tasks [32].

However, none of the works cited above considered the dynamic optimization of the data reduction strategy for multi-user goal-oriented communications, *jointly* with the global network resource management, under *prescribed performance guarantees*, as we do in this manuscript. Along this line, in [33] we proposed minimum-energy and maximum-accuracy resource allocation strategies for edge-assisted image classification tasks, in a *single user*/single server scenario, whereas in [34] we reported some preliminary results on the extension to the multi-user scenario, which we will further develop and investigate more thoroughly hereinafter.

Our contributions. The main contributions of this work concern the system architecture, the optimization strategies, and the simulation results. They can be summarized as follows:

A. System Architecture

Extending the preliminary strategies presented in [34], we consider a *multi-user* goal-oriented communication scenario, where *multiple* UEs may decide to offload their learning tasks to an ES (*or not*). Each user relies on a bank of source encoders, each one associated to a specific compression ratio, which dynamically compresses the data-units (DUs) to be transmitted to the ES, depending on the online system state. Specifically, exploiting convolutional encoders (CEs), i.e., the encoders of convolutional auto-encoders (CAE), as in [33], we improve their performance by a new training function. The ES, when requested, carries out multiple, user-independent, inference tasks, using a bank of convolutional classifiers (CCs), i.e., CNNs, each one matched to the CE used at the UE. The overall CE-CC structure is instrumental to *splitting the classification task between UE and ES*.

B. Optimization Strategies

We implement a *dynamical split* of the inference task, selecting, in each time slot, the most suitable pair of CE-CCs, within the bank of available (pre-trained) CE-CCs, depending on the channel state and on the online accuracy and performance. More specifically, resorting to Lyapunov optimization, we implement a multi-user *dynamical* goal-oriented source compression architecture that selects the CE-CC pair and allocates computational and communication resources, trading off energy consumption (including both UEs and ES), delay and classification accuracy. Hereinafter, we extend the preliminary results and optimization strategy shown in [34], by considering also a multi-user Maximum Accuracy strategy, with guaranteed (maximum) Delay bounds and Energy consumption (MADE). Furthermore, we let every UE able to decide whether to perform the inference task locally or to offload it to the ES, since there might be applications

where the UE hardware is capable of running the application locally, or it could be more convenient, for the overall resource management, to do that.

C. Simulation Scenarios

We investigate herein scenarios that were not analyzed in [34], where each UE has different service requirements and constraints. The wide set of possible scenarios, optimization strategies, and simulation results, significantly extends the results presented in [34], highlighting the effectiveness and flexibility of the proposed holistic resource management.

Outline. The paper is organized as follows. Section II illustrates the goal-oriented communication system and the related joint training procedure of both the CEs and the CCs for classification purposes. Section III describes the overall system model used in the formulation of the resource optimization strategies, which are then solved in Section IV exploiting stochastic Lyapunov optimization. In Section V we discuss our experimental results and, finally, in Section VI we draw some conclusions and highlight future research directions.

II. Classification Network and Training

This section describes the architecture employed to make parsimonious use of transmission energy and bandwidth. Specifically, we compress the UEs data-units (DUs) (i.e., the input of the learning task), before they are transmitted to the ES. The latter has to perform the learning task without sacrificing a prescribed target accuracy. As more deeply explained in [33], the Information Bottleneck (IB) [12] is a promising theoretical framework to meaningfully compress the data-source in a goal-oriented perspective. However, IB admits a closed form solution only when the associated statistics are discrete or Gaussian distributed [14], [15]. Thus, since in the multi-class image classification task we are focusing on, the Gaussian assumptions do not hold true and a meaningful definition of mutual information is problematic [35], we proposed in [33] a heuristic approximation of the IB that nicely fits with our goal-oriented strategy. Specifically, our approach is based on the deployment of a tunable data-compression at the UEs that is useful for the associated inference task at the ES. Without restriction of generality for the overall GOCs architecture and its resource management, we resort to banks of CEs to compress images at the UE side, according to a layer-by-layer max-pooling strategy. The CEs are coupled with CCs at the ES to perform the final decision, as summarized in Fig. 1 for a single UE.

As detailed in [33], a CE may be realized as:

- *Short-CE:* It resizes the images to the desired resolution by a single convolutional layer followed by a max-pooling layer.
- *Deep-CE:* It down-samples the images by multiple convolutional layers, each one followed by a max-pooling layer that halves the size of the (pseudo) image.

Note that our goal is to classify the images and not to reproduce them. Thus, for the CE-CCs compression and classification network shown in Fig. 1, we have to consider a

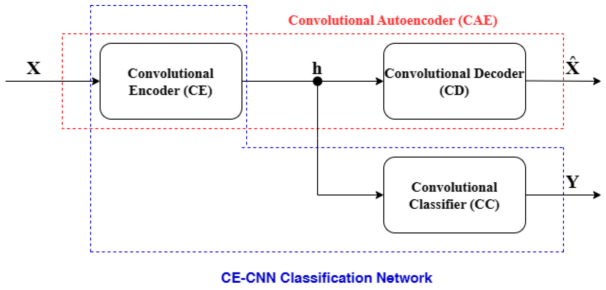


Fig. 1. Training scheme: the output of the CE h feeds both the ES classification CNN and a CD.

different learning cost function than those used for classical CAEs. Specifically, we resort to the following objective function

$$\underset{\theta,\phi}{\text{minimize}} \ \frac{1}{N_t}\sum_{n=1}^{N_t} L_{ce}\Big(Y_n, \hat{Y}_n, \phi, \theta\Big) + \lambda L_{mse}\Big(X_n, \hat{X}_n, \theta\Big), \tag{1}$$

where $L_{ce}(Y_n, \hat{Y}_n, \phi, \theta)$ is the cross-entropy loss, used in order to control the performance of the ES classification task, while $L_{mse}(X_n, \hat{X}_n, \theta)$ is the *Mean Squared Error* between the input and the reconstructed version $\hat{X}$ of the full CAE. Note that the cross-entropy loss in (1) is a proxy of the mutual information $I(h;Y)$ [36]. Thus, minimizing the cross-entropy, we maximize the $I(h;Y)$ for a fixed CE architecture (compression size) and this constitutes the link of the proposed approach with the IB principle. However, differently from what we did in [33], (1) considers also the output MSE of a Convolutional Decoder (CD), i.e., that part of the CAE that is typically used for image reconstruction. Actually, the presence in (1) of this (regularizing) MSE penalty term favours a meaningful feature extraction [37], which can improve the performance of the overall learning task, for proper values of the parameter λ. Anyway, note that the CD is taken into account only during the CE-CCs training, while it is not used for classification, as clarified by Fig. 1. Each (split) CE-CC couple has to be properly trained, possibly off-line, by a third party. Thus, although it would be interesting to analyze how to train the classification network by the same wireless edge-computing architecture we consider herein for classification, this is not the object of this manuscript and is left for future studies.

JPEG compression. Note that the CE, targeting good classification performance, compresses the images by a down-sampling principle, due to the max-pooling strategy at each layer. However, this design does not take into account the wireless communication between UEs and ES. Thus, while the size of the latent representation h of a CE output (see Fig. 1) may be optimal for a target classification accuracy, it could be still sub-optimal with respect to the file size of the compressed data-units, leading to huge costs in terms of transmission energy and (long) transmission time. This problem justifies the employment of a further zipping (compression) phase on h, before transmitting it to the ES, which will unzip it back to h at the CC input. Due to the nature of the classification task and the structure of the pseudo-images h extracted by the CE, we base this further compression at the UE on a JPEG codec, which proved to effectively reduce the file

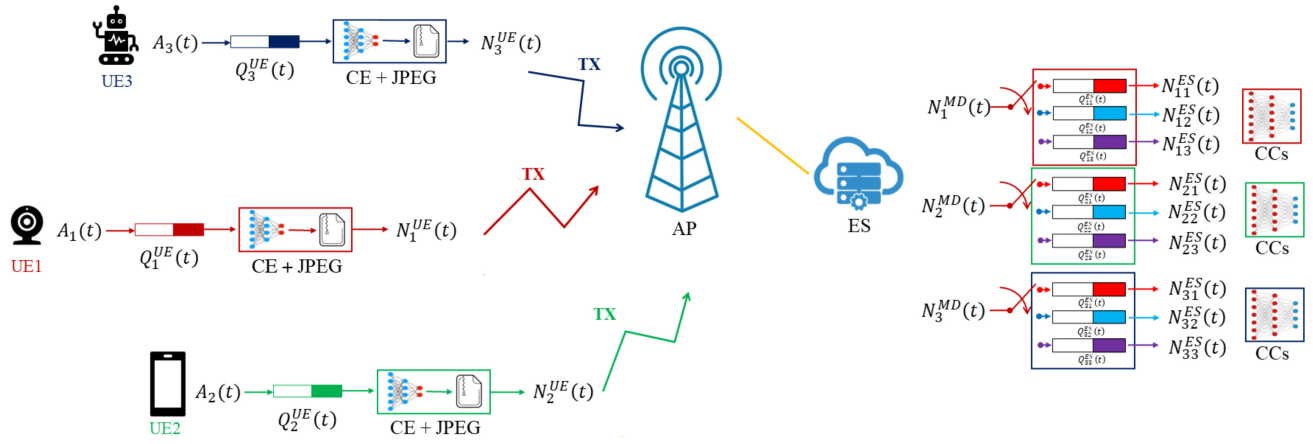


Fig. 2. Scenario: each UE dynamically employs its own set of CEs coupled with a proper set of CCs at the ES.

size of the data units, paying a reasonable price in terms of additional computational overhead from the UE perspective. The choice of JPEG is justified since it is a widely used zipping system, with a plethora of efficient implementations. Furthermore, despite its lossy nature, it has been proved that JPEG codecs do not significantly affect the classification performance of CNNs [38].

III. System Model

The considered goal-oriented scenario encompasses multiple devices (UEs), with limited computational and energy capabilities, which are connected through an Access Point (AP) to an ES with a larger amount of computing resources; an illustration is given in Fig. 2. To perform a generic learning task, for each UE connected to the network, the system handles three main phases: i) The UE buffers the Data Units (DUs), i.e., the images to be classified; ii) Depending on the specific offloading decision, which is affected by the system status, the DUs are either scheduled to be compressed and transmitted by the goal-oriented compression strategy proposed in Section II or, alternatively, to be processed locally; iii) The inference task takes place either at the UE- or ES-side, depending on the offloading decision.

The system evolves in a time-slotted fashion, where each time slot has a fixed duration τ. Therefore, we deal with discrete-time functions $f(t)$, where $t \in \mathbb{N}$ is an index for the t-th time-slot $[t\tau, (t+1)\tau[$. The aim of the resource optimization strategies for GOC is to guarantee a specific E2E (maximum) delay requirement, while optimizing either the system energy consumption or the learning accuracy. To this end, the proposed policies have to manage several resources. In particular, the k-th UE has to allocate its *transmission rate* $R_k(t)$ toward the ES, its *clock frequency* $f_k^d(t)$, employed to perform the data compression by a specific *compression factor* $\rho_k(t)$, and the offloading decision $d_k(t)$. As far as the ES is concerned, the main optimization variable is represented by the *clock frequency* $f_c(t)$, which has to be properly split among the learning tasks of the different users. This quantities represent the optimization variables of the objective functions we will define for the proposed resource management strategies. We are now ready to describe the models adopted for latency, energy and classification accuracy.

A. Latency Model

The system evolution over time is entirely described by a queuing system, as prescribed by the Lyapunov optimization framework [39]. In particular, for each user involved in the network, we define two kind of physical queues:

- A *computation/communication* queue at each UE, which collects the DUs, i.e., the images, generated by each device, which are waiting to be compressed and transmitted to the ES for classification.
- A separate *computation* queue at the ES side for any possible compression degree (e.g., CE) that the UEs may dynamically employ: thus, for each UE connected to the network, we have a different number of ES queues, depending on the CE compression degrees that are available. This design choice has been motivated in order to make the ES optimization problem computationally affordable, as we will clarify later.

We denote with K the total number of UEs connected to the network. The binary variable $d_k(t) \in \{0, 1\}$ models the decision to offload (or not) the learning task of the k-th device during the t-th time-slot. When any UE has to offload its learning task (i.e., $d_k(t) = 1$), we make the following assumptions that are instrumental to practically manage the optimization problem (see [33] for further details).

Assumption 1: The DUs in each UE queue have to be compressed and transmitted within the same time-slot. Indeed, during a given time-slot, it is impossible to optimally compress DUs that will be transmitted during one of the next time-slots, when the system could possibly experience different channel conditions, or different lengths of the ES/UEs queues, etc. Therefore, compression and transmission operations have to be done sequentially within the same time-slot.

Assumption 2: We assume that, while an UE is transmitting some DUs, it can also simultaneously compress other DUs.

The number of (compressed) DUs that would be possible to transmit during the t-th time-slot is expressed by

$$N_k^{tx}(t) = \left\lfloor \frac{\tau R_k(t)}{M(\rho_k(t))N(\rho_k(t))} \right\rfloor, \tag{2}$$

where $R_k(t)$ and $\rho_k(t)$ are the transmission rate and the compression factor,[1] respectively, selected for the k-th UE at time t; $M(\rho_k(t))$ is the DU's size for a certain compression factor $\rho_k(t)$, and $N(\rho_k(t))$ is the number of bits that are necessary (on average) to encode a pixel in the (zipped) pseudo-image h. To shorten the notation, we define also $W(\rho_k(t)) = M(\rho_k(t))N(\rho_k(t))$, which represents the average number of bits to store an image with a given $\rho_k(t)$. On the other hand, the number $N_k^c(t)$ of DUs that is possible to compress during the t-th time-slot by the k-th device is expressed by

$$N_k^c(t) = \left\lfloor \tau f_k^d(t) J_d(\rho_k(t)) \right\rfloor, \tag{3}$$

where $J_d(\rho_k(t))$ denotes the number of DUs compressed in a clock cycle C (which depends on the selected compression factor $\rho_k(t)$), and $f_k^d(t)$ denotes the device clock-frequency that has been chosen for the k-th UE, during the same time-slot. Recalling Assumption 1, all the DUs that are compressed within a time-slot have to be transmitted during the same time-slot, and all the transmitted DUs have to be first compressed. Thus, we need to use a transmission rate $R_k(t) \leq W(\rho(t)) f_k^d(t) J_d(\rho(t))$ which results in $N_k^{tx}(t) \leq N_k^c(t)$. Taking into account that, before the transmission could start, we need to wait a time equal to $1/(f_k^d(t) J_k^d(t))$ to compress the first DU, the actual number of DUs that can be offloaded by the k-th device during the t-th slot is expressed by

$$N_k^{off}(t) = \left\lfloor \frac{\tau - 1/\left(f_k^d(t) J_k^d(t)\right)}{W(\rho_k(t))/R_k(t)} \right\rfloor. \tag{4}$$

Plugging in (4) the inequality $N_k^{tx}(t) \leq N_k^c(t)$ we end-up with the following (integer) inequality

$$\left\lfloor \frac{\tau R_k(t)}{W(\rho_k(t))} \right\rfloor - 1 \leq N_k^{off}(t) \leq \left\lfloor \frac{\tau R_k(t)}{W(\rho_k(t))} \right\rfloor, \tag{5}$$

which will be useful in the next derivations.

Finally, similarly to (3), when the learning task is performed locally, the total number of DUs processed by the k-th UE is expressed by

$$N_k^L(t) = \left\lfloor \tau f_k^d(t) J_k^L(\rho_k(t)) \right\rfloor, \tag{6}$$

where $J_k^L(\rho_k(t))$ expresses the DUs that can be compressed by a factor $\rho_k(t)$ and successively classified in a clock-cycle by the UE hardware. Putting together (4) and (6), the number of DUs that can be processed by an UE, within a single time-slot, is expressed by

$$N_k^{UE}(t) = d_k(t) \cdot N_k^{off}(t) + (1 - d_k(t)) \cdot N_k^L(t). \tag{7}$$

The UE queue $Q_k^{UE}(t)$ is fed by the arrival of new DUs, and is drained either by the transmission of DUs to the ES, or by their local classification at the UE. Thus, it is characterized by the following evolution

$$Q_k^{UE}(t+1) = \max\left(0, Q_k^{UE}(t) - N_k^{UE}(t)\right) + A_k(t), \tag{8}$$

where $A_k(t)$ models the DUs arrival process, whose statistical properties are generally unknown.

At the ES, we employ L_k different queues for each UE, whose evolution is described by

$$Q_{ki}^{ES}(t+1) = \max\left(0, Q_{ki}^{ES}(t) - N_{ki}^{ES}(t)\right) + d_k(t) \cdot \min\left(N_k^{UE}(t), Q_k^{UE}(t)\right) \cdot \mathbb{1}_i\{\rho_k(t)\}, \tag{9}$$

i.e., a queue for each compression factor among the L_k in the set $\mathcal{S}_k = \{s_{ki}\}_{i=1,\dots,L_k}$, which represents the set of the compression factors employable by the k-th UE. These queues store the ES computation load, expressed in number of DUs, that is reserved for the k-th device. The term $\mathbb{1}_i\{\rho_k(t)\}$ in (9) is a shorthand for the indicator function $\mathbb{1}\{\rho_k(t) = s_{ki}\}$, which models the arrival of new DUs in the ES queue only if the UE have chosen the i-th compression factor. The term $N_{ki}^{ES}(t)$ in (9) denotes the number of DUs processed by the ES during the t-th time-slot, and it is expressed by

$$N_{ki}^{ES}(t) = \lfloor \tau f_{ki}^s(t) J_{ki}^s(t) \rfloor, \tag{10}$$

where $f_{ki}^s(t)$ is the ES clock-frequency assigned to the i-th queue (compression factor) of the k-th UE, during the t-th time slot.[2] The quantity $\frac{1}{J_{ki}^s(t)}$ in (10) is a conversion factor that maps the number of DUs received by the ES into the equivalent number of clock-cycles requested for their processing (e.g., classification).

To set-up our delay constraints, we need to define an overall queue that, for each device, takes into account the overall computational load at both the UE- and ES-side. Since we aim to respect an average latency constraint, as we will detail in the following, and taking in mind the ES can perform a parallel computation of multiple DUs, by means of (8) and (9), it makes sense to consider the average length of the parallel queues, which is expressed by

$$Q_k^{tot}(t) = Q_k^{UE}(t) + \sum_i^{L_k} p_{ki} Q_{ki}^{ES}(t), \tag{11}$$

where p_{ki} is the probability to employ the i-th compression factor in $\mathcal{S}_k$, which can be estimated by an online sample-mean.[3] By assuming a certain data arrival rate $\overline{A_k} = \mathbb{E}\left\{\frac{A_k(t)}{\tau}\right\}$, and exploiting the Little's Law [40], (11) allow us to model the average long-term delay, as expressed by

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{ \frac{Q_k^{tot}(t)}{\overline{A_k}} \right\}. \tag{12}$$

[1]Note that we denote with ρ the compression factor of the images along each dimension. The actual compression ratio scales with ρ^2.

[2]Having different queues for each compression factor is a design choice instrumental to obtain a mathematical dependence between N_{ki}^{ES} and f_{ki}^s, that is simpler than in [33], where we used a single queue. This way, the solution of the ES optimization problem becomes feasible also in a multi-user context, as we will clarify later.

[3]The $p_{k,i}$ are actually time-varying with the system state, which is also influenced by the *instantaneous* and adaptive resource management strategies we will end up with. The assumption here is that the stochastic resource management algorithms, which will exploit knowledge of the estimated $p_{k,i}$, will converge to a steady state where also the running sample mean estimate of the $p_{k,i}$ will converge. This fact has been verified by extensive simulation results.

For a latency constraint D_k^{avg}, we get a queue length constraint $Q_k^{avg} = D_k^{avg}\overline{A}_k$ and, consequently, we can equivalently formalize the latency constraint as a queue constraint by

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{ Q_k^{tot}(t)\} \leq Q_k^{avg}. \tag{13}$$

B. Energy Model

The energy model of our system involves three main components:

- *Transmission energy at the UEs*, requested to transmit the DUs to the ES in case of offloading decisions.
- *Computation energy at the UEs*, requested in order to either compress/encode the DUs to be transmitted, or to perform the learning task locally.
- *Computation energy at the ES*, requested to classify the DUs transmitted by the UEs that decide to offload the learning tasks.

For simplicity, assuming a capacity achieving transmission system, in a flat-fading wireless channel, the transmission power $p_k^{tx}(t)$ requested by the k-th UE can be inferred by the Shannon capacity [41]

$$R_k(t) = B_k \log_2\left(1 + \frac{p_k^{tx}(t)|h_k(t)|^2}{N_0 B_k}\right), \tag{14}$$

where $|h_k(t)|$ is the channel gain, N_0 denotes the noise power spectral density at the receiver side, and B_k is the bandwidth. Thus, by inverting (14), we obtain that the transmission energy spent by the k-th UE during the t-th time-slot depends on the rate $R_k(t)$ by

$$E_k^{tx}(t) = \tau p_k^{tx}(t) = \frac{\tau B_k N_0}{|h_k(t)|^2}\left(e^{\frac{R_k(t)ln(2)}{B_k}} - 1\right). \tag{15}$$

From the computation perspective, the ES's and UE's models are equivalent. Specifically, in order to estimate the energy consumption, we exploit the model in [42], which assumes a cubic dependence on the ES's and UE's clock-frequencies $f_s(t)$ and $f_k^d(t)$, as expressed by

$$E_k^d(t) = \tau\kappa_k^d f_k^d(t)^3 \text{ and } E_s(t) = \tau\kappa_s f_s(t)^3. \tag{16}$$

The constants κ_s and κ_k^d represent the effective switched capacitance [42] of ES and k-th UE processor, respectively. Thus, we quantify the system energy consumption during the t-th time-slot using the following weighted performance metric:

$$E_k^{tot}(t) = (1-\gamma)E_s(t) + \gamma\sum_{k=1}^{K}\delta_k\big(E_k^c(t) + E_k^{tx}(t)\big), \tag{17}$$

where the parameter γ is used to weight the UEs versus ES energy consumption, enabling tuning toward the implementation of an user-centric ($\gamma \to 1$) or a server-centric ($\gamma \to 0$) optimization strategy. Furthermore, the weights $\{\delta_k\}_{k=1}^K$ (with $\sum_{k=1}^K \delta_k = 1$) can be employed to assign different importance to the energy consumption of different users, providing an extra degree of flexibility to the resource optimization, depending on the needs of the operators, users, and service providers.

C. Accuracy Model

For the accuracy of the learning task of each UE, we resort to a *model-based* management strategy. This means that the accuracy for the k-th task can be cast in the optimization problem as a function $G_k(\rho_k(t))$ of the compression degree. This can be done in practice by employing a look-up table (LUT) (shown in Section V), where each entry is associated with a specific compression factor $\rho_k \in \mathcal{S}_k$.[4] This LUT stores the (average) classification accuracy of the k-th learning task, associated with each one of the CE-CC classifying chains that are available for the k-th UE. The values stored in this accuracy-LUT can be estimated off-line on meaningful test-sets, after each CE-CC structure has been properly trained, as described in the previous section. Thus, we can exploit the LUTs $G(\rho_k(t))$ to enforce an average accuracy constraint for each learning task, as expressed by

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{ G_k(\rho_k(t))\} \geq G_k^{avg}. \tag{18}$$

IV. Dynamic Resource Optimization for Multi-User Goal-Oriented Communications

On the basis of the delay, accuracy, and energy models presented in the previous section, we develop two resource optimization strategies: a multi-user Minimum-Energy with (maximum) Delay and Accuracy constraints (mu-MEDA), and a multi-user Maximum-Accuracy with (maximum) Delay and Energy consumption constraints (mu-MADE). In the sequel, we describe the problem formulation and the algorithmic solution for both strategies.

A. mu-MEDA: Multi-User Minimum-Energy With Delay and Accuracy Constraints

Following a system energy minimization perspective, the long-term optimization problem can be cast as follows:

$$\begin{aligned}
&\min_{\Phi(t)} \ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{E_{tot}(t)\} \\
&\text{s.t.} \ \ (a)\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{ Q_k^{tot}(t)\Big\} \leq Q_k^{avg}, \forall k \\
&\qquad (b)\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big\{ G\big(\rho_k(t)\big)\big\} \geq G_k^{avg}, \forall k \\
&\qquad (c)\ 0 \leq R_k(t) \leq R_{k,max}, \quad \forall k,t \\
&\qquad (d)\ \rho_k(t) \in \mathcal{S}_k, \quad f_s(t) \in \mathcal{F}_s, \quad f_k^d(t) \in \mathcal{F}_{d,k} \quad \forall k,t \\
&\qquad (e)\ \sum_{k=1}^{K}\sum_{i=1}^{L_k} f_{ki}^s(t) \leq f_s(t), \quad (f)\ f_{ki}^s(t) \geq 0 \quad \forall k,i,t \\
&\qquad (g)\ d_k(t) \in \{0,1\} \quad \forall k,t
\end{aligned} \tag{19}$$

[4] We modeled the relationship between the compression factor and the accuracy through a LUT, rather than by a formal analytical expression, because it is almost impossible to find a closed-form expression for this function in practice. Indeed, despite noticeable examples to theoretically formalize DNNs performance can be found in [43], [44], these approaches are based on Mutual Information, which is intractable to derive in closed-form in most of the practical cases.

where $\Phi(t) = [\{R_k(t), f_{ki}^s(t), f_k^d(t), \rho_k(t), d_k(t)\}_{k=1}^K, f_s(t)]$ contains all the optimization variables. The constraints in (19) have the following meaning: (*a*) the average queue length for the *k*-th UE must be lower than Q_k^{avg}, i.e., we are imposing a maximum average service delay equal to $D_{avg}^k = Q_{avg}^k/\overline{A_k}$ (cf. (13)); (*b*) the average classification accuracy for the *k*-th UE must be greater that G_{avg}^k; (*c*) the *k*-th UE transmission rate $R_k(t)$ must be smaller than the value $R_{k,max}(t)$, which is the maximum possible rate for the *k*-th device, inferred by (14), considering the maximum available transmission power $p_{k,max}^{tx}$; (*d*) specifies the discrete sets $\mathcal{F}_c$, $\mathcal{F}_{d,k}$ and $\mathcal{S}_k$ for the server frequencies set, the frequencies set for the *k*-th UE, and the set of the possible compression factors respectively; the constraints $(e) - (f)$ state that the sum of the clock frequencies $f_{ki}^s(t)$ that the (edge) server allocates for all the queues assigned to each user, must be lower than the total ES clock-frequency chosen for the *t*-th time slot, and that each clock-frequency must be obviously grater than 0; finally, (*g*) represents the binary constraints on the set of the opportunistic offloading decisions variables of each UE. Problem (19) is complicated due to the lack of knowledge of the statistics of the radio channels and data arrivals, which would be necessary to compute the expected values in (19). To tackle this issue, we resort to Lyapunov stochastic optimization arguments [39], which solve the long term problem (19) by casting it to a sequence of instantaneous optimization problems, which can be solved in a per-slot fashion. According to such an optimization framework [39], we start associating a *virtual* queue to each one of the long-term constraints (*a*) and (*b*). These virtual queues evolve according to

$$\begin{aligned} Z_k(t+1) &= \max\big(0, Z_k(t) + \mu_k\big(Q_k^{tot}(t+1) - Q_k^{avg}\big)\big) \\ Y_k(t+1) &= \max\big(0, Y_k(t) + \nu_k\big(G_k^{avg} - G_k(t)\big)\big), \end{aligned} \tag{20}$$

where μ_k and ν_k are step-sizes that control the convergence speed of the algorithm. This way, it is possible to prove that respecting the long term constraints $(a) - (b)$ is equivalent to guarantee the mean-rate stability of the virtual queues in (20) [39]. To this end, we define the actual Lyapunov function *L*(*t*), as the sum of the squares of all the (virtual and physical) queues

$$L(t) = \sum_{k=1}^K Z_k(t)^2 + \sum_{k=1}^K Y_k(t)^2. \tag{21}$$

Defining $\Theta(t) = \big[\{Z_k(t)\}_{k=1}^K, \{Y_k(t)\}_{k=1}^K\big]$, we obtain the associated conditional *Lyapunov drift*

$$\Delta(\Theta(t)) = \mathbb{E}\{L(t+1) - L(t) | \Theta(t)\}, \tag{22}$$

whose minimization corresponds to the stabilization of the virtual queues, but it does note take into account the objective function (i.e., the system energy consumption). Thus, in order to trade-off system stability and energy consumption, the Lyapunov Drift is augmented with a term dependent on the system energy, to obtain the so-called *Lyapunov Drift plus Penalty* function

$$\Delta_p(\Theta(t)) = \Delta(\Theta(t)) + V\mathbb{E}\{E_{tot}(t)\}. \tag{23}$$

By increasing the value of the parameter *V* we give more importance to the objective function rather than to the queues stability, thus pushing the solution toward optimality while still guaranteeing the stability of the system, i.e., respecting the long-term constraints. In particular, [39] proved that, as the parameter *V* increases, the optimal solution of (19) is asymptotically reached. Following stochastic optimization arguments [39], we proceed minimizing an upper bound of the Lyapunov Drift plus penalty function in (23) (derived in the Appendix), ending up with the instantaneous optimization problem in (24), where, since the optimization variables affect only the terms N_k^{UE}, N_{ki}^{ES} and G_k, we neglect all the terms which do not depend on them. Note moreover that in the following we omit the time index *t* to simplify the notation.

$$\begin{aligned} \min_{\Phi} \quad & VE_{tot} + \sum_{k=1}^K \Bigg[L_k N_k^{UE} \mu_k^2 \Bigg(\sum_{i=1}^{L_k} \mathbb{1}_i\{\rho_k\} p_{ki} Q_{ki}^{ES} - Q_k^{UE} \Bigg) \\ & - L_k \mu_k^2 \sum_{i=1}^{L_k} p_{ki} Q_{ki}^{ES} N_{ki}^{ES} + \mu_k Z_k \Bigg(\max\Big(0, Q_k^{UE} - N_k^{UE}\Big) \\ & + \sum_{i=1}^{L_k} \max\Big(0, p_{ki} Q_{ki}^{ES} - N_{ki}^{ES}\Big) \Bigg) - \nu_k Y_k G_k(\rho_k) \Bigg] \\ \text{s.t.} \quad & 0 \le R_k \le R_{k,max}, \quad \rho_k \in \mathcal{S}_k, \quad f_s \in \mathcal{F}_s, \quad f_k^d \in \mathcal{F}_{d,k} \\ & \sum_{k=1}^K \sum_{i=1}^{L_k} f_{ki}^s \le f_s, \quad f_{ki}^s \ge 0, \quad \forall k, i. \end{aligned} \tag{24}$$

Since the UEs energy-consumption terms in the cost function of problem (24) depend only (and separately for each UE) on the UEs optimization variables $\{\Phi_{d,k}\}_{k=1}^K = \{[R_k, f_k^d, \rho_k, d_k]\}_{k=1}^K$, we can optimize this part of the cost function separately at each UE. Note that our design choice to assign at the ES separate computation queues for each UE offloaded task, lets us completely decouple the optimization problem and separately handle the UE and ES resource optimization. Furthermore, as already pointed out in footnote 2, the use of multiple queues for each compression factor ρ_{ki}, thanks to (11), makes by (10) the problem linear with respect to f_{ki}, up to the $\lfloor . \rfloor$ operator. Consequently, Problem (24) is separable and solvable for each compression factor, as described in the following.

1) UE Sub-Problem: For the *k*-th device, at each time slot *t*, we have to solve the following optimization problem

$$\begin{aligned} \min_{\Phi_{d,k}} \quad & L_k N_k^{UE} \mu_k^2 \Bigg(\sum_{i=1}^{L_k} \mathbb{1}_i\{\rho_k\} p_{ki} Q_{ki}^{ES} - Q_k^{UE} \Bigg) \\ & + \mu_k Z_k \max\Big(0, Q_k^{UE} - N_k^{UE}\Big) - \nu_k Y_k G_k(\rho_k) \\ & + V\gamma\delta_k\big(E_k^{tx} + E_k^c\big) \\ \text{s.t.} \quad & 0 \le R_k \le R_{k,max}, \quad \rho_k \in \mathcal{S}_k, \quad f_k^d \in \mathcal{F}_{d,k}, \\ & d_k \in \{0, 1\}. \end{aligned} \tag{25}$$

Depending on the value of the offloading decision variable d_k we can optimize the other variables employing two different strategies. If $d_k = 1$, we have to allocate both the

transmission rate R_k to transmit the DUs to the ES, and the UE clock-frequency f_k^d and compression factor ρ_k to perform compression. Otherwise, if $d_k = 0$ we need only to allocate f_k^d and ρ_k to perform the learning task locally. We remark that we assume, although this is not mandatory, that the UE employs also locally the same (bank of) CE-CC classification chains we designed for the GOC scheme, thus fairly offering to the UEs the same flexibility of classification accuracy and energy consumption that could be exploited by the ES solution. Other choices, or a fixed structure of the classifier at the UE, would obviously have an impact on the offloading decisions by the optimal resource management and, consequently, on the energy-delay-accuracy tradeoffs.

Coming to the solution of the problem, when $d_k = 1$ we handle the $\min(\cdot)$ in (4) by adding the following constraint on the transmission rate of the k-th user

$$0 \leq R_k \leq R_{k,max}^+, \; R_{k,max}^+ = \min\left\{ R_{k,max}, \frac{Q_k^{UE} W(\rho_k)}{\tau} \right\}. \tag{26}$$

This way, according to Assumptions 1 and 2, and taking in mind we cannot compress more DUs that we can transmit, we select a data-rate that is bounded by the minimum between the *maximum achievable rate* $R_{k,max}$ (computed plugging the maximum power p_k^{tx} in the Shannon capacity (14)), and the *draining rate* $Q_k^{UE} W(\rho_k)/\tau$ that is capable to empty the transmission queue (and lets remove the $\max(\cdot)$). By considering that $x - 1 \leq \lfloor x \rfloor \leq x$, we can also remove the $\lfloor \cdot \rfloor$ in (4). Therefore, using the definition of the indicator function, for any *fixed* compression factor $\rho_{ki} \in \mathcal{S}_k$, we end up with the following optimization problem

$$\min_{\Phi_{d,k}} \; -\frac{Q_{ki}^{TX} \tau R_k}{W(\rho_k)} + \frac{\tau V \gamma \delta_k B_k N_0}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} + \tau V \gamma \delta_k \kappa \left(f_k^d\right)^3 - \nu_k Y_k G_k(\rho_k) \tag{27}$$

$$\text{s.t.} \;\; 0 \leq R_k \leq R_{k,max}^+, \quad f_k^d \in \mathcal{F}_{d,k}, \tag{28}$$

where $Q_{ki}^{TX} = L_k \mu_k^2 (Q_k^{UE} - p_{ki} Q_{ki}^{ES}) + \mu_k Z_k$. This is a mixed-integer optimization problem. However, in practice, the sets $\mathcal{F}_{d,k}$ and S_k have a quite low cardinality and, as detailed below, the solution can be rapidly found by an exhaustive search. Indeed, for any fixed couple of compression factor $\rho_k \in \mathcal{S}_k$ and computation frequency $f_k^d \in \mathcal{F}_{d,k}$, the optimization problem is convex with respect to the data rate R_k, whose optimal value can be found in closed form by *duality theory* through the Lagrangian

$$\mathcal{L} = -\frac{\tau Q_{ki}^{TX} R_k}{M(\rho_k) N(\rho_k)} + \frac{\tau V \gamma \delta_k N_0 B_k}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} + \tau V \gamma \delta_k \kappa \left(f_k^d\right)^3 - \nu_k Y_k G_k(\rho_k) - \alpha R_k + \beta \left(R_k - R_{k,max}^+\right), \tag{29}$$

where α and β are the Lagrangian multipliers. Note that, if $Q_{ki}^{TX} \leq 0$, the second term monotonically increases with the rate, and the Lagrangian is minimum for $R_k = 0$.

Otherwise, when $Q_{ki}^{TX} > 0$ we can solve the optimization problem by imposing the following KKT conditions [45]

$$(a) \; \frac{\partial \mathcal{L}}{\partial R_k} = -\frac{Q_{ki}^{TX} \tau}{W(\rho_k)} + \frac{\tau V \gamma \delta_k ln(2) N_0 B_k}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} - \alpha + \beta = 0$$

$$(b) \; 0 \leq R_k \leq R_{k,max}^+, \; (c) \;\; \alpha \geq 0, \; (d) \;\; \beta \geq 0$$

$$(e) \;\; \alpha R_k = 0, \; (f) \;\; \beta \left(R_k - R_{k,max}^+\right) = 0. \tag{30}$$

Solving the KKT conditions we can compute the optimal rate $R_k^*(\rho_k, f_k^d)$, by the following expression

$$R_k^* = \left[\frac{B_k}{\ln(2)} \ln\left(\frac{Q_{ki}^{TX} h_k^2}{W(\rho_k) V \gamma \delta_k \ln(2) N_0} \right) \right]_0^{R_{max}^+} \times \mathbb{1}\left(Q_{ki}^{TX} > 0\right), \tag{31}$$

which gives us the closed form expression for the optimal rate for any fixed compression factor ρ_k and clock frequency f_k^d, of the k-th user. Thus, as anticipated, to select the best clock frequency f_k^{d*}, and compression factor ρ_k^*, we can proceed by an exhaustive search, thanks to the limited cardinality of $\mathcal{F}_{d,k}$ and $\mathcal{S}_k$. Summarising, for a potential offloading $(d_k = 1)$, we compute the optimal rate and clock frequency f_k^d for each possible compression factor ρ_k, and then, at every time slot, we select the triple $T_k^* = (R_k^*, f_k^{d*}, \rho_k^*)$ that gives the lowest energy cost. Otherwise, for a potential classification at the UE $(d_k = 0)$, the transmission rate to the ES would be $R_k = 0$ and we need to optimize only the clock-frequency for each possible compression factor, thus obtaining the optimal pair $P_k^* = (f_k^{d*}, \rho_k^*)$ that minimizes the UE's energy consumption. The overall optimal solution of the UE's optimization problem, which includes the decision to offload or not the learning task, is finally given by choosing between the pairs $(d_k = 1, T_k^*)$ and $(d_k = 0, P_k^*)$, as the one that leads to the minimum value of the UE's energy cost function.

2) ES Sub-Problem: From the ES perspective, for each UE we have to manage multiple computing queues, each one associated to a specific compression factor that has been used by the specific UE: in the following, we denote with Q_{ki}^{ES} the i-th ES computing queue for the k-th UE. It clearly makes sense to constrain the fraction f_{ki}^s (of the total ES's computing frequency f_s) reserved to the i-th queue of the k-th user, to be lower than what would be necessary to completely drain the same queue within a time-slot, as expressed by

$$f_{ki}^s \leq \min\left(f_s, \frac{Q_{ki}^{ES}}{\tau J_{ki}^s} \right). \tag{32}$$

This way, we can remove the terms $\max(0, Q_{ki}^{ES} - N_{ki}^{ES})$ from the sum in (24) and, consequently, we can rewrite the ES's resource allocation problem as

$$\min_{\Phi_s} \; -\sum_{k=1}^{K} \sum_{i=1}^{L_k} \tau Q_{ki}^{comp} J_{ki}^s f_{ki}^s + \tau V (1-\gamma) \kappa f_s^3 \tag{33}$$

$$\text{s.t.} \;\; 0 \leq f_{ki}^s \leq \min\left(f_s, \frac{Q_{ki}^{ES}}{\tau J_{ki}^s} \right) \;\; \forall k, i$$

$$\sum_{k=1}^{K} \sum_{i=1}^{L} f_{ki}^s \leq f_s, \quad f_s \in \mathcal{F}_s,$$

where $\Phi_s = [\{f_{ki}^s\}_{i=1,\ldots,L_k, k=1,\ldots,K}, f_s]$, and $Q_{ki}^{comp} = L_k \mu_k^2 Q_{ki}^{ES} + \mu_k Z_k$. Although the problem is a mixed-integer optimization one, for any fixed ES's clock frequency f_s, it

boils down to the classical (fractional) knapsack problem [46]. Consequently, the optimal solution is obtained by a greedy algorithm, which consists in ordering the queues by their weights $(Q_{ki}^{comp} J_{ki}^{s})$ in descending order, and then assigning the clock frequency to the queue as $\min(\overline{\phi}, \frac{Q_{ki}^{ES}}{\tau J_{ki}^{s}})$, where $\overline{\phi}$ is the remaining part of the ES's clock frequency $f_c(t)$. Consequently, due to the limited cardinality of the ES's clock-frequency set $\mathcal{F}_s$, also in this case we can exhaustively solve the problem for all the server clock frequencies $f_s \in \mathcal{F}_s$, thus obtaining the set of possible solutions $\{(f_{ki}^s, f_s)\}_{f_s \in \mathcal{F}_s}$ and then choose the one associated with the minimum ES's cost in (33).

B. mu-MADE: Multi-User Maximum-Accuracy With Delay and Energy Constraints

An alternative resource allocation, targeting a Maximum-Accuracy, can be formulated as

$$\min_{\Phi(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\{\sum_{k=1}^{K} -G_k(t)\right\} \tag{34}$$

$$\text{s.t. } (a)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{Q_k^{tot}(t)\} \le Q_k^{avg} \quad \forall k$$

$$(b)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\{E_k^d(t)\right\} \le E_k^{d,avg} \quad \forall k$$

$$(c)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{E_s(t)\} \le E_s^{avg}$$

$$(d)\ 0 \le R_k(t) \le R_{k,max} \quad \forall k,t$$

$$(e)\ \rho_k(t) \in \mathcal{S}_k,\ f_s(t) \in \mathcal{F}_s,\ f_k^d(t) \in \mathcal{F}_{d,k} \quad \forall k,t$$

$$(f)\ \sum_{k=1}^{K}\sum_{i=1}^{L_k} f_{ki}^s(t) \le f_s(t)$$

$$(g)\ f_{ki}^s(t) \ge 0 \quad \forall k,i,t$$

$$(h)\ d_k(t) \in \{0,1\} \quad \forall k,t$$

where $\Phi(t) = [\{R_k(t), f_{ki}^s(t), f_k^d(t), \rho_k(t), d_k(t)\}, f_s(t)]$, for $k = 1,\dots,K$, and $i = 1,\dots,L_k$ contains all the optimization variables. The constraints in (19) have the following meaning: (a) the average queue length for the k-th UE must be lower than Q_k^{avg}, i.e., we are imposing a maximum average service delay equal to $D_{avg}^k = Q_{avg}^k/\overline{A_k}$ (cf. (13)); (b) the average energy consumption for the k-th UE must be lower than $E_{d,avg}^k$; (c) the average ES's energy consumption must be lower than E_{avg}^s; (d)-(h) have the same meaning of (c)-(g) in (19).

Proceeding similarly to the mu-MEDA strategy, in order to manage the long-term energy constraints (b) and (c), in addition to the virtual queue $Z_k(t)$ defined in (20) to manage (a), we need to define the virtual queues

$$\begin{aligned} S_k(t+1) &= \max\Big(0, S_k(t) + \lambda_k\Big(E_k^d(t+1) - E_k^{d,avg}\Big)\Big) \\ O(t+1) &= \max(0, O_k(t) + \eta(O(t+1) - E_s^{avg})), \end{aligned} \tag{35}$$

where $\{\lambda_k\}_{k=1}^K$ and η are the step-sizes used to control the convergence speed of the algorithm. By the definition of the virtual queues, in this case the Lyapunov Function becomes

$$L(t) = \sum_{k=1}^{K}\Big[S_k(t)^2 + Z_k(t)^2\Big] + O(t)^2 \tag{36}$$

and, consequently, given $\Theta(t) = [\{S_k(t), Z_k(t)\}_{k=1}^K, O(t)]$, we derive the following expression for the Lyapunov *drift-plus-penalty* function

$$\Delta_p(t) = \mathbb{E}\{L(t+1) - L(t)|\Theta(t)\} - V\mathbb{E}\left\{\sum_{k=1}^{K} G_k(t)\right\} \tag{37}$$

As detailed in the Appendix, we end up with the following optimization problem

$$\begin{aligned} \min_{\Phi}\ & \sum_{k=1}^{K}\Bigg[L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES} - Q_k^{UE}\right) \\ & + \mu_k Z_k\Bigg(\max\Big(0, Q_k^{UE} - N_k^{UE}\Big) \\ & + \sum_{i=1}^{L_k}\max\Big(0, p_{ki}Q_{ki}^{ES} - N_{ki}^{ES}\Big)\Bigg) \\ & + \lambda_k S_k E_k^d - L_k\mu_k^2\sum_{i=1}^{L_k} p_{ki}Q_{ki}^{ES}N_{ki}^{ES}\Bigg] \\ & + \eta O E_k^s - V\sum_{k=1}^{K} G_k \\ \text{s.t. } & 0 \le R_k \le R_{k,max},\ \rho_k \in \mathcal{S}_k,\ f_s \in \mathcal{F}_s,\ f_k^d \in \mathcal{F}_{d,k}\ \forall k \\ & \sum_{k=1}^{K}\sum_{i=1}^{L_k} f_{ki}^s \le f_s,\ f_{ki}^s \ge 0 \end{aligned} \tag{38}$$

Exploiting again the decoupling of the problem, which is granted by our proposed design to separately handle the queues for any specific UE and any specific compression factor, we end-up also in this case with distinct instantaneous optimization problems, one at each UE, and a single one at the ES.

1) UE Sub-Problem: As far as the k-th UE is concerned, we get the following optimization problem formulation

$$\begin{aligned} \min_{\Phi_{d,k}}\ & L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\Big\{\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES}Q_k^{UE}\right) \\ & + \mu_k Z_k\max\Big(0, Q_k^{UE}N_k^{UE}\Big) \\ & - \lambda_k S_k E_k^d - V G_k(\rho_k) \\ \text{s.t. } & 0 \le R_k \le R_{k,max} \\ & \rho_k \in \mathcal{S}_k, f_k^d \in \mathcal{F}_{d,k}, d_k \in \{0,1\}, \end{aligned} \tag{39}$$

where $\Phi_{d,k} = [R_k, f_k^d, \rho_k, d_k]$, for $k = 1,\dots,K$. The resolution strategy is quite similar to the previous case, when we minimized the energy consumption: if an UE would decide to offload its task $(d_k = 1)$, we need to allocate the optimal transmission rate R_k for any fixed compression factor ρ_k and

TABLE I
DEEP-CE PARAMETERS

ρ	$G(\rho)$ [%]	$J_k^d(\rho)$ $[\frac{DU}{C}]$	$J_k^L(\rho)$ $[\frac{DU}{C}]$
2	97.3	1.44×10^{-7}	8.35×10^{-8}
4	96.5	1.26×10^{-7}	9.04×10^{-8}
8	93.4	1.16×10^{-7}	8.90×10^{-8}
16	91.8	1.07×10^{-7}	8.73×10^{-8}
32	83.0	1.35×10^{-7}	1.06×10^{-7}
64	67.0	1.32×10^{-7}	1.09×10^{-7}

device clock frequency f_k^d. Also in this case we can obtain the optimal rate $R_k^*(\rho_k, f_k^d)$ in closed form, as expressed by

$$R_k^* = \left[\frac{B_k}{\ln(2)} \ln\left(\frac{Q_{ki}^{TX} h_k^2}{W(\rho_k)\lambda_k S_k \ln(2) N_0} \right) \right]_0^{R_{max}^+} \times \mathbb{1}\left(Q_{ki}^{TX} > 0 \right). \tag{40}$$

Thus, for a possible offloading decision ($d_k = 1$) we compute by (40) the optimal data transmission rate R_k^* for each $\rho_k \in S_k$ and $f_k^d \in F_{k,d}$, and we select the optimal triple $T_k^* = (R_k^*, f_k^{d*}, \rho_k^*)$ that minimizes the cost function in (39). Conversely, in order to evaluate the minimum cost of a local learning task at the k-th UE ($d_k = 0$), we just need to exhaustively search for the pair $P_k^* = (f_k^{d*}, \rho_k^*)$ that would optimize the accuracy under the prescribed constraints. Finally, depending on which one of the two optimal allocation strategies guarantees the best accuracy, we decide to offload ($d_k = 1$), or not ($d_k = 0$), the k-th user task, using the associated optimal allocation strategy T_k^*, or P_k^*, respectively.

2) ES Sub-Problem: From the ES perspective, the optimization problem is similar to the mu-MEDA, except for small differences in the cost function, and is expressed by

$$\min_{\Phi_s} \; -\sum_{k=1}^{K} \sum_{i=1}^{L_k} \tau Q_{ki}^{ES} J_{ki}^s f_{ki}^s + \eta O \kappa \tau f_s^3 \tag{41}$$

$$\text{s.t. } 0 \le f_{ki}^s \le \min\left(f_s, \frac{Q_{ki}^{ES}}{\tau J_{ki}^s} \right), \quad \forall k, i$$

$$\sum_{k=1}^{K} \sum_{i=1}^{L} f_{ki}^s \le f_s, \quad f_s \in \mathcal{F}_s,$$

where $\Phi_s = [f_{ki}^s, f_s]$, and can be solved likewise the mu-MEDA formulation.

V. SIMULATION RESULTS

In this section, we present the simulation results we obtained by the two optimization strategies we proposed and solved. Tables I-II report the values of the accuracy $G_k(\rho)$, the data-units $J_k^d(\rho)$ that can be compressed (and zipped by JPEG) in a clock-cycle by the k-th UE, when it decides to offload the classification, and the data-units $J_k^L(\rho)$ that can be compressed and classified locally in a clock-cycle by the same UE. Table III reports the data-units $J_s(\rho)$ that can be classified in a clock-cycle at the ES, as well as the image-size $M(\rho)$ and the average number of bits/pixel $N(\rho)$ that are shared by both the short- and deep-CE, when using JPEG.

TABLE II
SHORT-CE PARAMETERS

ρ	$G(\rho)$ [%]	$J_d(\rho)$ $[\frac{DU}{C}]$	$J_k^L(\rho)$ $[\frac{DU}{C}]$
2	97.3	1.44×10^{-7}	8.35×10^{-8}
4	95.8	1.68×10^{-7}	1.10×10^{-7}
8	91.5	1.88×10^{-7}	1.26×10^{-7}
16	91.3	1.95×10^{-7}	1.38×10^{-7}
32	77	2.25×10^{-7}	1.55×10^{-7}
64	50.0	2.25×10^{-7}	1.65×10^{-7}

TABLE III
COMMON PARAMETERS

ρ	$M(\rho)$ $[px]$	$N(\rho)$ $[\frac{bits}{px}]$	$J_s(\rho)$ $[\frac{DU}{C}]$
2	128x128x3	1.08	1.2×10^{-7}
4	64x64x3	2.27	2.17×10^{-7}
8	32x32x3	4.72	2.87×10^{-7}
16	16x16x3	9.06	3.57×10^{-7}
32	8x8x3	8	5×10^{-7}
64	4x4x3	8	6.25×10^{-7}

TABLE IV
CHANNEL TYPE

Ch. Type	D $[m]$	B $[kHz]$	f_0 $[GHz]$	σ_0^2
A	50	2500	6	1.06×10^{-10}
B	500	2500	9	2.72×10^{-14}

We assumed a flat-fading channel, whose statistical characterization is based on the *Clarke's autocorrelation function* [47]. We considered two operating scenarios, summarized in Table IV, and we accordingly set the time-slot duration to $\tau = 50ms$, which corresponds to the channel coherence time. The parameter σ_0^2 models the wireless channel power path-loss and it has been computed by considering the *Alpha-Beta-Gamma* model [48]. In a first set of simulations we considered a scenario with $K = 5$ UEs connected to the network. Although this is not strictly necessary, we assumed that the devices of all the UEs share the same computation frequency set $\mathcal{F}_d = \{0.1, 0.2, \ldots, 0.9, 1\} \times 1.4GHz$, while the server computation frequency set is $\mathcal{F}_s = \{0.1, 0.2, \ldots, 0.9, 1\} \times 4.5GHz$. Finally, for simplicity, we considered an effective switched capacitance $\kappa = 1.097 \times 10^{-27} [\frac{s}{cycles}]^3$ for all the UEs and for the ES. We underline that all the simulation results have been obtained at convergence of the tested strategies [39].

A. Goal-Oriented Compression Results

For simplicity, all the UEs were assigned the same image classification task, based on the German Traffic Sign Recognition Benchmarks (GTSRB) [49] dataset. This dataset includes 1213 pictures of German road signals, divided in 43 different classes. The dataset has been split in a 80% training set, composed of 970 images, and 20% test set, composed of 243 images. During the data loading phase, all the all the images have been normalized to a size of 256x256, and converted to a 3-channel image (one channel for each RGB color), such that the initial size of each data-unit, is 256x256x3. Although this is not strictly necessary, we assumed that all the UEs share the same bank of CE-CC classification networks, e.g., the compression factors ρ_k assume values on the same fixed set $\mathcal{S} = \{2, 4, 8, 16, 32, 64\}$. In order to shade light on

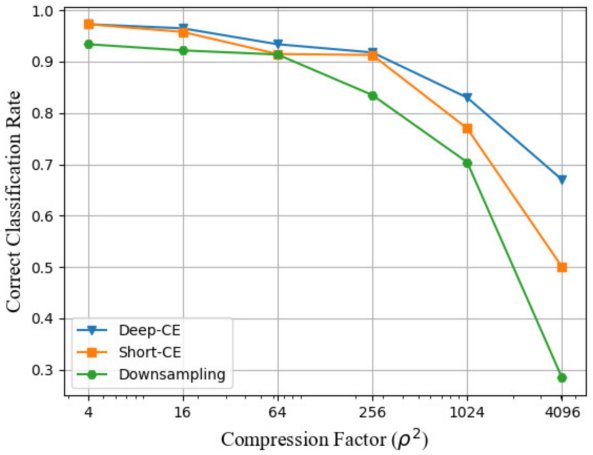


Fig. 3. Classification accuracy comparison.

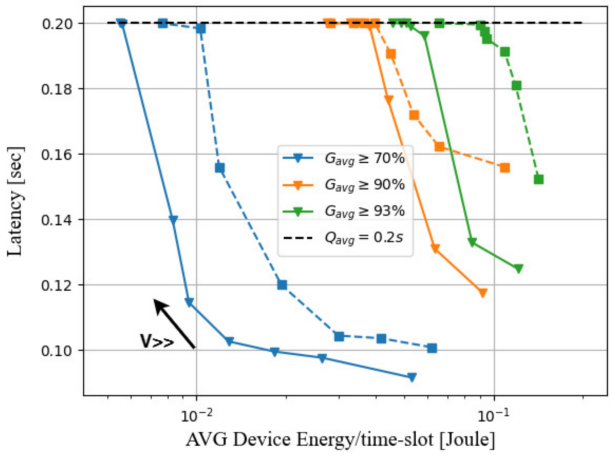


Fig. 4. UE Energy/Latency trade-off. CE (solid) vs down-sampling (dashed).

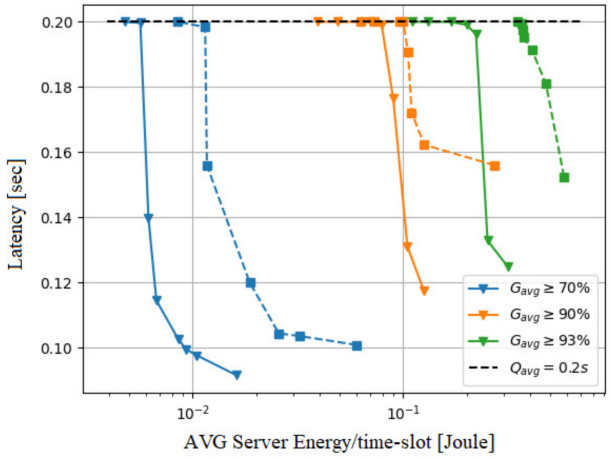


Fig. 5. ES Energy/Latency trade-off. CE (solid) vs down-sampling (dashed).

the performance obtained by the proposed resource managements, we find useful to show in Fig. 3 the average accuracy on the test-set associated to different compressive architectures: i) Deep-CE, ii) Short-CE, iii) Down-sampling with anti-aliasing pre-filter. As expected, the accuracy $G(\rho)$ has a monotone decreasing behavior with respect to the compression factor, for all the models. The deep-CE has always the best performances even if, for lower compression factors (up to 16), the differences with the Short-CE are almost negligible. In contrast, for the highest ones (i.e., 32, 64) there is there is a clear advantage in using the deep-CE. For compression factor $\rho = 64$ we get output tensors with a size of 4x4x3=48 pixels: despite (pseudo) images of this size have clearly undergone a heavy transformation, the deep-CE still allows the ES's CC to classify them with a 67% accuracy, which is still a remarkable performance for a 43-class 43-class classification task. Conversely, for this compression factor neither the down-sampling strategy nor the short-CE, allow a meaningful classification. The price to be paid for an increased accuracy of the deep-CE is the increase of the computation energy and processing delay (as summarized in Tables I-II) that we trade by our resource management policies.

B. mu-MEDA Results

First of all, we tested the mu-MEDA strategy comparing the CE (short and deep) with the down-sampling compression strategy in channel scenario *B*, reported in Table IV. We set the same *latency* constraint $D_k^{avg} = Q_k^{avg}/\overline{A}_k = 0.20\ s$, for all the UEs. We considered a task arrival process with $\overline{A}_k = 2DU/slot$, and we forced the UEs to always offload the classification task to the ES, without any opportunistic strategy (i.e., $d_k(t) = 1,\ \forall k, t$).

Each trade-off curve in Figs. 4 and 5 is associated to a different accuracy constraint, while they all respect the same latency constraint, which is highlighted by a dashed horizontal line in the plot. Each curve is obtained by evaluating the solution (at convergence) of the resource optimization problem, for several different values of the trade-off parameter V in (23). Specifically, by increasing V we end-up to solutions characterized by a lower energy consumption and a higher latency and, as indicated by the black arrow on the figures, we move from the bottom-right to the top-left corner of the trade-off plots, which correspond to the desired optimal solutions on the borders of the feasibility regions. Fig. 4 shows that, from the UE's perspective, there is a clear advantage on employing the CE compression strategy, since we end-up to solutions characterized by a lower (computational and transmission) energy consumption, while satisfying the same latency and accuracy constraints. This depends on the fact that *channel-B* is characterized by a huge attenuation: thus, since the CE compression strategy allows to satisfy the same accuracy constraint transmitting smaller DUs with respect to classical down-sampling, this allows to reduce the transmission energy expenditure considerably, without spending too much in extra computational energy for CE-based compression at the UE. Actually, the proposed dynamical, goal-oriented, compression strategy leads also to a lower ES's energy computational expenditure, as witnessed from Fig. 5. Indeed, also the classification of smaller DUs is cheaper from a computational and energetic perspective.

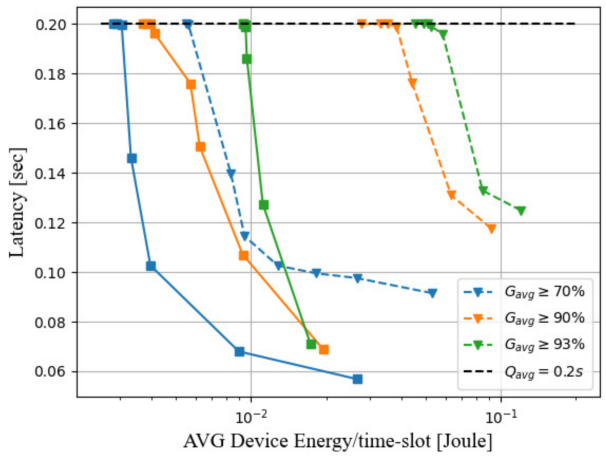


Fig. 6. UE's energy/latency trade-off. Opportunistic offloading (solid) vs only offloading strategy (dashed).

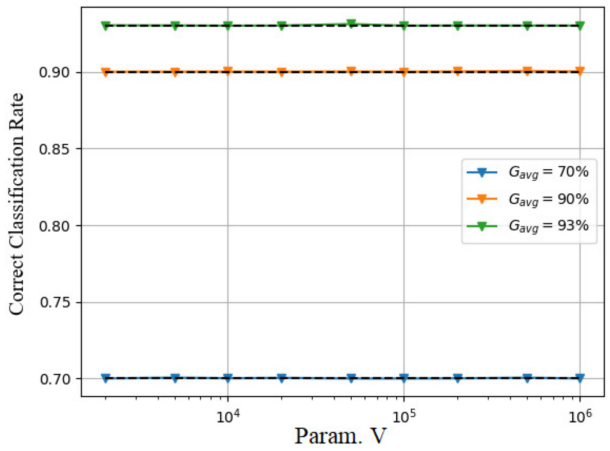


Fig. 7. Average Accuracy vs V with opportunistic offloading at convergence.

C. Opportunistic Offloading

We compared the previous scenario, where UEs always offload decision tasks to the ES, with the opportunistic offloading strategy where UEs can also decide to perform classification locally, by the same CE-CC classification architecture. Specifically, two out of five UEs are connected to the ES by the channel in *scenario A* of Table IV, while the other ones by the channel in *scenario B*. The opportunistic offloading strategy ends up to a dynamical resource optimization that is characterized by a significant lower UE energy expenditure with respect to the *always offload* strategy, still satisfying both the accuracy and latency constraints, as shown by Figs. 6-7, where clearly all the solid curves are on the left, e.g., with a lower energy expenditure, with respect to the dashed curves of the pure offloading strategy. Fig. 8 shows the histogram of the offloading decisions for each UE, for a (minimum) accuracy constraint $G_{avg} = 70\%$ and a trade-off parameter $V = 1 \times 10^6$. As expected, since the UE-0 and UE-3 experience good channel conditions, they decide to offload more frequently than the other devices, whose Channel-B requests much higher transmission power to allocate rates to the UEs

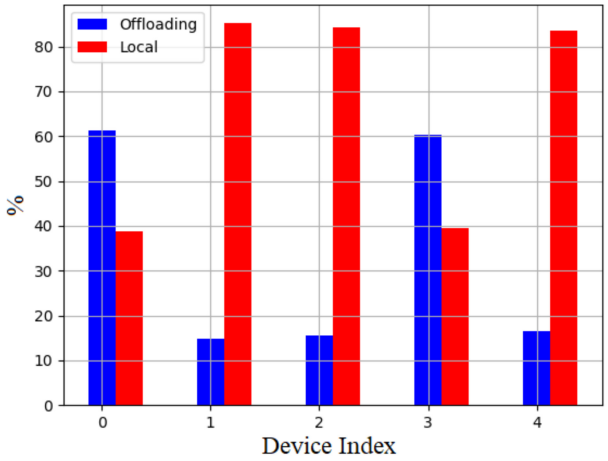


Fig. 8. % of Offloading ($G_k^{avg} = 70\% \forall k,\ V = 1 \times 10^6$).

TABLE V
SIMULATION SCENARIOS FOR EACH UE

UE	Ch. Type	$\kappa[\frac{s}{cycles}]^3$
0	**A**	$10 \times \kappa_0$
1	**A**	$20 \times \kappa_0$
2	**B**	$30 \times \kappa_0$

and, sometimes, it may be also unfeasible to respect either the accuracy or the delay constraint, or both.

D. Comparison With Static Allocation Strategies

A key strength of the proposed approach is the joint *dynamic* optimization of transmission&computational resources, together with the optimal dynamic selection of the classification architecture used to perform the task. Thus, we compare the proposed multi-user optimization strategy with:

- A *Fixed-Accuracy* optimization strategy, where we optimize both the computational and the transmission resources at the UE-side, by keeping fixed a single CE-CC classification architecture. This approach is quite similar to the one presented in [6].
- A *Hybrid* static/dynamic optimization strategy, where, inspired by [50], we fix the transmission rate R on the basis of the average channel conditions, while we dynamically optimize the CE-CC architecture, as well as the computational resources at the UEs. The transmission rate R is fixed as the minimum one that guarantees the stability of the UE queue. This rate can be computed through the capacity for flat-fading Rayleigh channels [51, eq. (9)], and it fixes also the transmission power.

In this case we considered a scenario with $K = 3$ UEs, each one experiencing different channel conditions and computational efficiency, as summarized in Table V. We set an arrival task with $\overline{A} = 2DU/slot$, and we imposed the same accuracy and latency constraints for all the UEs to $G_k^{avg} = 92\%$ and $D_k^{avg} = 0.2s$, respectively. Thus, for the *Fixed-Accuracy* optimization strategy, we considered the

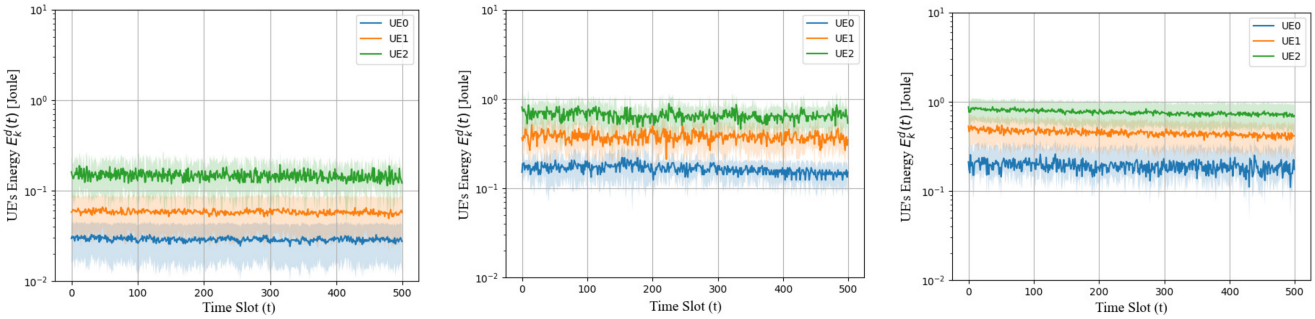


Fig. 9. Instantaneous UEs consumption for the dynamic optimization (a), fixed accuracy (b) and fixed rate (c).

short-CE with $\rho_k = 8$ as the unique learning model, which according to Table II is capable to grant the requested average classification performance with a fairly moderate computational energy. Fig. 9 shows that employing a fully dynamic optimization strategy leads to solution characterized by a lower UE energy consumption. As expected UE-0 and UE-2 reach the lowest and highest energy consumption, respectively, given their computational and channel conditions summarized in Tab. V. It is clear that, for all the UEs, our optimization strategy allow to reach the lowest energy consumption, thus confirming the effectiveness to *jointly and dynamically* optimize the transmission/computation resources as well as the learning architecture (i.e., the pair of CE-CC) to be employed, depending on the instantaneous system conditions.

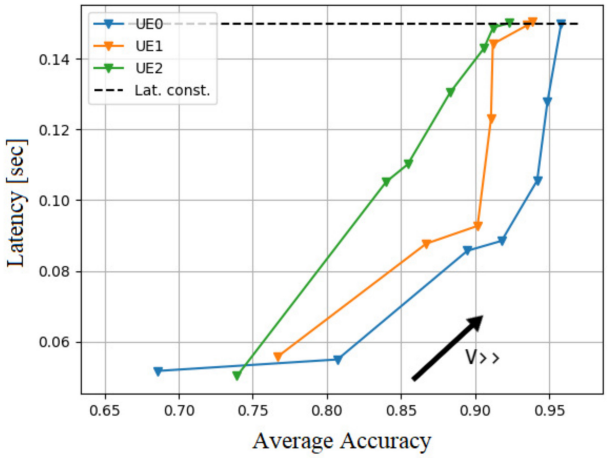


Fig. 10. Accuracy vs Latency trade-off.

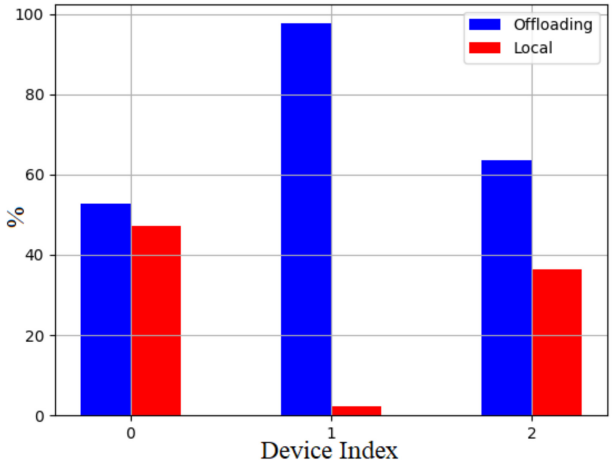


Fig. 11. Offloading histograms ($V = 1 \times 10^5$).

E. mu-MADE Results

We tested the mu-MADE optimization strategy considering a scenario with $K = 3$ UEs, each one characterized by different channel and computational conditions. In particular, we considered an effective switched capacitance $\kappa_0 = 1.097 \times 10^{-27}[\frac{s}{cycles}]^3$ for the ES, and higher values for the UEs, in order to simulate a lower energetic efficiency. The UE energy constraint has been set to $E_k^{avg} = 128 \times 10^{-3} J$. Table V summarizes the different conditions for the devices considered in the simulation, where we employed, concurrently, both Deep- and the Short-CE. We remark that UE-0 experiences both good channel conditions and computational efficiency: this means that it has the maximum degree of flexibility on the management of the opportunistic offloading. UE-1 is characterized by the same channel conditions of UE-0, with a lower computational efficiency, while UE-2 operates with both a bad channel and a low computational energy efficiency.

The curves shown in Fig. 10 represent the accuracy-latency trade-off: by increasing the parameter V of (37), we end up with solutions with higher accuracy and latency, moving on the curves from bottom-left to top-right corner, where we get the desired optimal solutions at the boundary of the decision region. Specifically, Fig. 10 shows that UE-0 (i.e., the UE with the best computational & channel conditions) gets the highest accuracy, while widely satisfying the latency constraint. We note a similar behaviour for UE-1 and UE-2, with a higher degree of latency for UE-2 (i.e., the device that works in the worst conditions). Finally, we report in Fig. 11 the histogram of the offloading decisions for each UE. Given its favorable channel and computational energy efficiency, we have a balanced situation for UE-0, since it has the highest flexibility to choose if offloading computations, or not. On the other hand, UE-1 mostly performs offloading, since the transmission of DUs in a channel with fairly low attenuation allows to mitigate the burden due to the low computational energy efficiency. Finally UE-2, although it has a much worse channel, it offloads

more DUs than UE-0 s due to its much higher computational inefficiency.

VI. Conclusion and Future Directions

In this work we implemented a goal-oriented compression architecture based on CEs, which is exploited by two distinct dynamic optimization strategies in order to either minimize the energy consumption or to maximize the learning accuracy in a multi-user scenario, where the UEs can opportunistically decide whether and when to offload the computations toward the ES. The extensive simulation results confirmed the effectiveness and the flexibility of the proposed approaches in different scenarios. However, we remark that the proposed goal-oriented communication architecture, and the associated resource management strategy, could exploit also classification or learning-oriented compression strategies, that may be different from the CE-based solutions presented herein. Future research directions include the extension to multi-server scenarios, cooperative learning tasks (e.g., Federated Learning), as well as to explicitly take into account also the battery level of each UE, which may be equipped by some energy harvesting mechanism or batteries recharge plan.

Appendix

Mathematical Derivations for MU-mEDA

Two Lemmas in [39] are useful to solve the proposed resource optimization strategies.

Lemma 1: Given a queue that evolves according to $X(t+1) = max(0, X(t) + x(t+1) - \overline{x})$, by defining $\Delta_x = \frac{X(t+1)^2 - X(t)^2}{2}$, it is always true that $\Delta_x \leq \frac{(x(t+1)-\overline{x})^2}{2} + X(t)x(t+1) - X(t)\overline{x}$.

Lemma 2: The following inequality holds true:

$$(\max(0, Q-b)+A)^2 \leq Q^2 + A^2 + b^2 + 2Q(A-b).$$

Employing Lemma 1, and recalling that, given $x \in \mathbb{R}^k$, $(\sum_{k=1}^K x_k)^2 \leq K \sum_{k=1}^K x_k^2$, for the Latency Virtual Queue $Z_k(t)$ we have

$$\begin{aligned}\Delta_{z_k}(t) &\leq \frac{\mu_k^2 \big(Q_k^{tot}(t+1) - Q_k^{avg}\big)^2}{2} \\ &\quad + \mu_k Z_k(t)\big(Q_k^{tot}(t+1) - Q_k^{avg}\big) \\ &\leq \frac{\mu_k^2 L_k}{2}\left[Q_k^{UE}(t+1)^2 + \sum_{i=1}^{L_k} p_i Q_{ki}^{ES}(t+1)^2\right] \\ &\quad + \mu_k Z_k(t)\big(Q_k^{tot}(t+1) - Q_k^{avg}\big) + \frac{\mu_k^2 L_k}{2}\big(Q_k^{avg}\big)^2,\end{aligned}$$

Now, recalling (8), (9) and using Lemma 2 we can derive the following inequality

$$\begin{aligned}\Delta_{z_k}(t) \leq \frac{\mu_k^2 L_k}{2}\Big\{ & Q_k^{UE}(t)^2 + M_k^{UE} + 2Q_k^{UE}(t) \\ & \times \Big(A_k(t) - N_k^{UE}(t)\Big) \\ & + \sum_{i=1}^{L_k}\Big[p_i Q_{ki}^{ES}(t)^2 + M_k^{ES} \\ & + 2p_i Q_{ki}^{ES}(t)\Big(\hat{A}_{ki}(t) - N_{ki}^{ES}(t)\Big)\Big]\Big\} \\ & + \mu_k Z_k(t)\Big(Q_k^{tot}(t+1) - Q_k^{avg}\Big) + \frac{\mu_k^2 L_k}{2}\big(Q_k^{avg}\big)^2\end{aligned}$$

where $\hat{A}_{ki}(t) = \mathbb{1}\{\rho_k(t) = s_{ki}\}N_k^{UE}(t)$, $M_k^{UE} = A_{k,max}^2 + N_{kdev,max}^2$ and $M_{ik}^{ES} = A_{ki,max}^2 + N_{ki,max}^2$. The same derivations presented in [9] can be applied to the accuracy virtual queue, thus obtaining an upper-bound for $\Delta_{y_k}(t)$. Putting together the derived instantaneous upper-bounds we end up to the optimization problem presented in Section IV-A.

References

[1] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[3] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[4] S. Wang et al., "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. IEEE INFOCOM*, 2018, pp. 63–71.

[5] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[6] M. Merluzzi, P. Di Lorenzo, and S. Barbarossa, "Wireless edge machine learning: Resource allocation and trade-offs," *IEEE Access*, vol. 9, pp. 45377–45398, 2021.

[7] 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.

[8] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[9] M. Merluzzi, P. Di Lorenzo, and S. Barbarossa, "Dynamic resource allocation for wireless edge machine learning with latency and accuracy guarantees," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 9036–9040.

[10] M. Merluzzi, C. Battiloro, P. Di Lorenzo, and E. C. Strinati, "Energy-efficient classification at the wireless edge with reliability guarantees," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2022, pp. 109–114.

[11] M. Merluzzi, A. Martino, F. Costanzo, P. Di Lorenzo, and S. Barbarossa, "Dynamic ensemble inference at the edge," in *Proc. IEEE Global Commun. Conf.*, 2021, pp. 1–6.

[12] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*.

[13] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[14] F. Pezone, S. Barbarossa, and P. Di Lorenzo, "Goal-oriented communication for edge learning based on the information bottleneck," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 8832–8836.

[15] G. Chechik, A. Globerson, N. Tishby, Y. Weiss, and P. Dayan, "Information bottleneck for Gaussian variables," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 165–188, 2005.

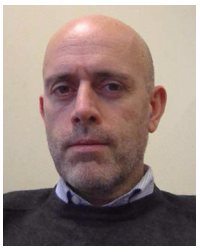
[16] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[17] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 73–87, Jan. 2023.

[18] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[19] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for Internet of Things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.
[20] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.
[21] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gunduz, "Deepjscc-q: Constellation constrained deep joint source-channel coding," 2022, *arXiv:2206.08100*.
[22] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided deep joint source channel coding for wireless Multipath fading channels," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 584–599, Jun. 2022.
[23] J. Dai et al., "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 2300–2316, Aug. 2022.
[24] J. Ballé et al., "Nonlinear transform coding," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 2, pp. 339–353, Feb. 2021.
[25] M. K. Farshbafan, W. Saad, and M. Debbah, "Common language for goal-oriented semantic communications: A curriculum learning framework," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 1710–1715.
[26] M. K. Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," 2022, *arXiv:2204.10429*.
[27] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.
[28] X. Peng et al., "A robust deep learning enabled semantic communication system for text," 2022, *arXiv:2206.02596*.
[29] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep source-channel coding for sentence semantic transmission with HARQ," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5225–5240, Aug. 2022.
[30] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Aug. 2022.
[31] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 553–557, Mar. 2022.
[32] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," 2021, *arXiv:2107.11190*.
[33] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Adaptive resource optimization for edge inference with goal-oriented communications," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, pp. 1–34, 2022.
[34] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Dynamic resource allocation for multi-user goal-oriented communications at the wireless edge," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, 2022, pp. 697–701.
[35] K. G. Larkin, "Reflections on shannon information: In search of a natural information-entropy for images," 2016, *arXiv:1609.01117*.
[36] M. Boudiaf et al., "A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 548–564.
[37] L. Le, A. Patterson, and M. White, "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 107–117.
[38] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. 8th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, 2016, pp. 1–6.
[39] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synth. Lect. Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.
[40] J. D. Little, "A proof for the queuing formula: L= λ W," *Oper. Res.*, vol. 9, no. 3, pp. 383–387, 1961.
[41] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[42] T. Burd and R. Brodersen, "Processor design for portable systems," *J. VLSI Signal Process.*, vol. 13, pp. 203–221, Nov. 1996.
[43] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2015, pp. 1–5.
[44] A. M. Saxe et al., "On the information bottleneck theory of deep learning," *J. Statist. Mech. Theory Exp.*, vol. 2019, no. 12, 2019, Art. no. 124020.
[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[46] M. Assi and R. A. Haraty, "A survey of the knapsack problem," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, 2018, pp. 1–6.
[47] A. F. Molisch, *Statistical Description Of The Wireless Channel*. New York, NY, USA: Wiley, 2011.
[48] S. Sun et al., "Propagation path loss models for 5G urban micro- and macro-cellular scenarios," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, 2016, pp. 1–6.
[49] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 1453–1460.
[50] C. Liu, C. Guo, Y. Yang, and N. Jiang, "Adaptable semantic compression and resource allocation for task-oriented communications," 2022, *arXiv:2204.08910*.
[51] J. Li, A. Bose, and Y. Q. Zhao, "Rayleigh flat fading channels' capacity," in *Proc. 3rd Annu. Commun. Netw. Services Res. Conf. (CNSR)*, 2005, pp. 214–217.

Francesco Binucci (Graduate Student Member, IEEE) received the bachelor's and M.Sc. degrees in computer engineering from the University of Perugia, Perugia, Italy, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in industrial and information engineering with the Department of Engineering. He is currently a Research Collaborator with the Consorzio Nazionale Interuniversitario per le Telecomunicazioni. His research interests include resource management for wireless communications, edge machine learning, signal processing theory and methods, and data science.



Paolo Banelli (Member, IEEE) received the Laurea degree (cum laude) in electronics engineering and the Ph.D. degree in telecommunications from the University of Perugia, Perugia, Italy, in 1993 and 1998, respectively, where he has been a Full Professor with the Department of Engineering since 2019, an Associate Professor since 2005, and an Assistant Professor since 1998. He was a Visiting Researcher with the University of Minnesota, Minneapolis, MN, USA, in 2001, and a Visiting Professor with Stony Brook University, Stony Brook, NY, USA, from 2019 to 2020. His current research interests include signal processing theory and methods, wireless communications and edge intelligence, goal-oriented communications, graph signal processing, and distributed learning. In 2009, he was the General Co-Chair of the IEEE International Symposium on Signal Processing Advances for Wireless Communications. He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2013 to 2016 and *EURASIP Journal on Advances in Signal Processing* from 2013 to 2020, and has been serving as an Associate Editor for the IEEE OPEN JOURNAL OF SIGNAL PROCESSING since 2020. He was a member of the IEEE Signal Processing for Communications and Networking Technical Committee from 2011 to 2013.

Paolo Di Lorenzo (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the Sapienza University of Rome, Rome, Italy, in 2008 and 2012, respectively, where he is currently an Associate Professor with the Department of Information Engineering, Electronics, and Telecommunications. In 2010, he was a Visiting Researcher with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA, USA. From May 2015 to February 2018, he was an Assistant Professor with the Department of Engineering, University of Perugia, Perugia, Italy. He has participated in the FP7 European research projects FREEDOM, on femtocell networks; SIMTISYS, on moving target detection and imaging using a constellation of satellites; and TROPIC, on communication, computation, and storage over collaborative femtocells. He is a Principal Investigator of the research unit (CNIT-Sapienza) in the H2020 European Project RISE 6G. His research interests include signal processing theory and methods, distributed optimization, wireless edge intelligence, goal-oriented and semantic communications, and graph signal processing. He was a recipient of the three Best Student Paper Awards, respectively, at IEEE SPAWC10, EURASIP EUSIPCO11, and IEEE CAMSAP11. He was also a recipient of the 2012 GTTI (Italian National Group on Telecommunications and Information Theory) Award for the Best Ph.D. thesis. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.

Sergio Barbarossa (Fellow, IEEE) received the M.S. and Ph.D. degrees in EE from the Sapienza University of Rome, where he is currently a Full Professor and a Senior Research Fellow of the Sapienza School of Advanced Studies. He has held visiting positions with the Environmental Research Institute of Michigan in 1988, the University of Virginia in 1995 and 1997, and the University of Minnesota in 1999. He has been the Scientific Coordinator of several EU projects on wireless sensor networks, small cell networks, distributed mobile cloud computing, and edge computing in 5G networks. He is currently leading a national project on edge learning and he is involved in two H2020 European projects on 5G networks for Industry 4.0 and on reconfigurable intelligent surfaces. His current research interests are in the area of mobile-edge computing and machine learning, graph signal processing, and distributed optimization. He received the IEEE Best Paper Award from the IEEE Signal Processing Society in 2000, 2014, and 2020. He received the Technical Achievements Award from the EURASIP Society in 2010. He coauthored the papers that received the Best Student Paper Award at ICASSP 2006, Signal Processing Advances in Wireless Communications (SPAWC) 2010, EUSIPCO 2011, and CAMSAP 2011. He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1998 to 2000 and from 2004 to 2006, the *IEEE Signal Processing Magazine*, and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS. He has been the General Chairman of the IEEE Workshop on SPAWC in 2003 and the Technical Co-Chair of SPAWC in 2013. He has been the Guest Editor for Special Issues on the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *EURASIP Journal of Applied Signal Processing*, *EURASIP Journal on Wireless Communications and Networking*, the *IEEE Signal Processing Magazine*, and the IEEE SELECTED TOPICS ON SIGNAL PROCESSING. From 1997 to 2003, he was a member of the IEEE Technical Committee for Signal Processing in Communications. He is an EURASIP Fellow. He has been an IEEE Distinguished Lecturer.

Open Access funding provided by 'Università degli Studi di Perugia' within the CRUI CARE Agreement