Goal-oriented Spectrum Sharing: Trading Edge Inference Power for Data Streaming Performance

Mattia Merluzzi[1] and Miltiadis C. Filippou[2]
[1]CEA-Leti, Université Grenoble Alpes, F-38000 Grenoble, France
[2]WINGS ICT Solutions, 17121 Athens, Greece
e-mail: mattia.merluzzi@cea.fr, mfilippou@wings-ict-solutions.eu

***Abstract*—We study the problem of spectrum sharing between goal-oriented (GO) and legacy data-oriented (DO) systems. For the former, data quality and representation is no longer optimized based on classical communication key performance indicators, but rather configured on the fly to achieve the goal of communication with the least resource overhead. This paradigm can be followed to flexibly adapt wireless and in-network artificial intelligence operations across different nodes (e.g., access points, users, sensors or actuators) to data traffic, channel conditions, energy availability and distributed computing capabilities. In this paper, we argue and demonstrate that computing and learning/inference operation performance strongly affect lower layers, calling for a real cross-layer optimization that encompasses physical and computation resource orchestration, up to the application level. Focusing on a communication channel shared among a GO and a DO user, we define a *goal-effective achievable rate region (GEARR),* to assess the maximum data rate attainable by the latter, subject to goal achievement guarantees for the former. Finally, we propose a cross-layer dynamic resource orchestration able to reach the boundaries of the GEARR, under different goal-effectiveness and compute resource consumption constraints.**

***Index Terms*—Goal-oriented semantic communications, adaptive computation, resource allocation, spectrum sharing.**

I. Introduction

Semantic and goal-oriented (GO) communication aims at dynamically tailoring data representation and transmission, as guided by specific application needs [1]. Within the scope of this promising paradigm for 6G, communication performance requirements are *adapted to achieve the communication goal*, rather than set a priori and ossified. Wireless resource sharing between semantic, GO and legacy data-oriented (DO) services has several implications on network architectural design, along with radio and computing resource deployment and orchestration. It comes with challenges in terms of system backward compatibility, but also opportunities for more efficient spectrum use, thanks to the extraction of relevant information, also possibly exploiting the much lower application data quality that can be tolerated during transmission for some tasks. An indicative example of such tasks refers to the ones involving advanced Artificial Intelligence (AI)-based processing, including computer vision models dedicated to image classification or object detection that exhibit substantial robustness to noise (or, bit-level errors). Semantic data extraction and processing as part of a GO communication system setup need computing resources to be capillary available at end devices and edge nodes (e.g., edge server (ES)). Such edge resources facilitate the achievement of low two-way latency requirements, energy consumption reduction, data privacy and security, with data being kept as local as possible. Further, these resources assist with extracting relevant information, thus overcoming wireless signaling drawbacks (e.g., bit-level errors). This capability magnifies with increased computing capacity, and represents an opportunity for more efficient spectrum sharing.

Related works. A few works have already focused on spectrum coexistence between semantic, GO and DO systems. In [2], the authors propose a semi-non-orthogonal multiple access (NOMA) scheme for a two-users downlink communication, to improve the achievable rate for a DO user. However, the authors focus solely on the conveyed semantics, while overlooking the goal of communication. A similar approach is proposed in [3] for uplink communication, considering different multiple access schemes, to characterize the trade-off between semantic user rate and DO user rate. Again, the communication goal is limited to correctly receiving message meaning. In [4], GO communication is introduced in the problem, with a scheme that proposes to learn an adaptation of goal-achieving communication quality metrics to DO user interference, to allow a GO user to achieve its goal, i.e., confident and timely inference. None of these works proposes goal-aware adaptation of computing resources to the quality of received data in case of unfavorable channel conditions.

Contribution. We tackle this heterogeneous service coexistence problem from an interference perspective, to show how radio and computation resource domains are tightly related. Going beyond previous works, we propose to incorporate computing resource awareness and inference model availability into the resource orchestration policy. We define the concept of *goal-effective achievable rate region (GEARR)*, and propose a dynamic method to jointly control DO user transmit power, inference model selection for the GO user, proactive packet drops and computation resources, to explore its boundaries.

Notation: in the remainder of the paper, bold lower case letters denote vectors, while calligraphic letters denote sets. Also, given a random variable X, its long-term average is always denoted as $\bar{X}$, and defined as

$$\bar{X} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{X(t)\}. \tag{1}$$

This work has been supported by the SNS JU project 6G-GOALS under the EU's Horizon program Grant Agreement No 101139232, and by the ANR under the France 2030 program, grant "NF-NAI: ANR-22-PEFT-0003". For Miltiadis C. Filippou this work initiated while being with Nokia Strategy & Technology, 81541 Munich, Germany. Fig. 1 has been partly designed using resources from Flaticon.com

II. System model

The system under investigation is composed of two users, namely a GO user (UE_g), and a DO user, (UE_d), both served by the same Access Point (AP) in uplink, on the same frequency resources. The AP is equipped with N antennas and a computing node that is embedded with a set $\mathcal{L}$ of pre-trained AI models, ready to output inference results for UE_g. The system is illustrated in Fig. 1. The role of the buffer and the *control valve* at UE_d will be clarified in Sec. III. Denoting by $\mathbf{h}_{g(d)}(t) \in \mathbb{C}^{N\times 1}$ the complex channel for $\text{UE}_{g(d)}$ at a given time instant t, we can write the (instantaneous) Signal-to-Noise-plus-Interference-Ratio ($\text{SINR}_{d(g)}$) as

$$\text{SINR}_{g(d)}(t) = \frac{|\mathbf{w}_{g(d)}^H(t)\mathbf{h}_{g(d)}(t)|^2 p_{\text{tx},g(d)}(t)}{|\mathbf{w}_{g(d)}^H(t)\mathbf{h}_{d(g)}(t)|^2 p_{\text{tx},d(g)}(t) + \sigma_n^2}, \quad (2)$$

with $p_{\text{tx},g(d)}$ and σ_n^2 denoting the transmission power of UE_g (UE_d) and the noise power, respectively; whereas $\mathbf{w}_{g(d)}$ is the AP combining vector for the GO/DO user. We assume that the AP has instantaneous channel knowledge and applies a signal reception technique, e.g., Maximum Ratio Combining (MRC).

A. Key Performance Indicators (KPIs)

In this paper, we are interested in wireless performance of both the GO and DO user. For UE_g, we need to consider the following communication KPIs: i) communication delay and ii) communication reliability. For the latter metric, we use the Bit Error Rate (BER), which affects inference performance, as detailed in the sequel. Further, for UE_g, we consider computing delay as impacting inference timeliness. For UE_d we consider the data rate as the dominant communication KPI to support the running application, which is, however, rather insensitive to the relevance of communicated data. The objective is to assess the performance of the interference channel in terms of ***goal-effective achievable rate regions***.

1) Goal-oriented user: wireless delay, BER and inference

We assume the AP to dynamically select, at each time t, a modulation order $M(t) \in \mathcal{M}$ for UE_g, with an M-QAM constellation. Given $M(t)$, the wireless communication delay to upload a new inference input data batch reads as $D_{\text{tx}}(t) = N_b(t)/R_g(t)$, where $N_b(t)$ is the number of bits encoding one data batch, $R_g(t) = W\log_2(M(t))$ is the UE_g data rate and W is the available uplink bandwidth. Note that N_b could evolve over time thanks to possibly different source compression schemes that depend on the available resources and the specific context on the fly. However, in this paper we keep it fixed over time. The BER $P_b(t)$ depends on $\text{SINR}_g(t)$ and $M(t)$, and for an uncoded modulation is given by [5]:

$$P_b(t) = \frac{4}{\log_2(M(t))}\left(1 - \frac{1}{\sqrt{M(t)}}\right) Q\left(\sqrt{\frac{3\cdot \text{SINR}_g(t)}{M(t)-1}}\right).$$

2) Computing aspects: delay and inference KPIs

In this work, computing only concerns the GO user. Also, we assume the ES to be embarked with a set of $\mathcal{L} = \{1,\dots,L\}$ inference models (e.g., AI models) capable of addressing the GO user's task, each one with different computational complexity, thus different performance and robustness to data noise. Every model l is characterized by a tuple $(\omega_l, \Gamma_l(P_b))$, where ω_l is the number of Floating-Point-Operations (FLOPs) needed to run one inference instance (i.e., inference on one data batch)[1], and $\Gamma_l(P_b)$ is the model *reliability* (in this case, inference correctness probability, or, accuracy), which we define as the probability of issuing a correct inference result under a BER P_b. Instantaneously, the overall reliability depends on the AI inference model, the BER, and the specific data batch. The instantaneous inference correctness information given a specific input data batch is usually not retrievable during operation, since the ground truth might not be available (e.g., for an object detection or classification task). However, $\Gamma_l(P_b)$ can be estimated a priori on a validation set in the case of a supervised task (the effectiveness of this strategy will be shown in Sec. IV). In this paper, we assume the validation set to be drawn from the distribution of the test set, leaving issues related to distribution shifts to future investigations. Otherwise, other metrics, such as the entropy at the output of the classifier can be employed, as in [4]. During slot t, the computing delay depends on the selected inference model $l(t)$ and on the computing capacity allocated by the ES. Denoting the latter by $F(t)$ (measured in Floating Point Operations Per Second - FLOPS), the computing delay reads as $D_{\text{comp}}(t) = \omega_{l(t)}/F(t)$. Finally, the total delay for UE_g, including communication and computing delays, is $D_{\text{tot}}(t) = D_{\text{tx}}(t) + D_{\text{comp}}(t)$, where we assume that the inference output transmission in the downlink to be of negligible duration and over a different channel. Finally, the overall delay and the reliability of the employed inference model contribute to the achievement or the failure of the goal, which, in this case, corresponds to ***correctly classifying data within a deadline***, as clarified in Sec. II-B.

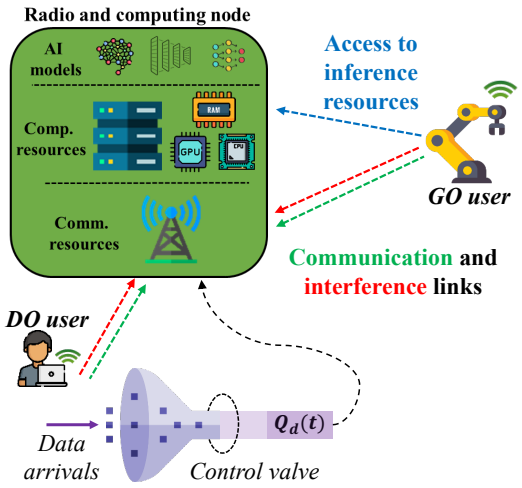


Fig. 1: Reference scenario

[1]Note that ω_l is an average value that does not take into account sample-specific computational cost. We leave this for future work.

3) Data-oriented user KPIs

As detailed in Sec. III, the objective of the DO user is to maximize its *average sustained data arrival rate* (i.e., the data arrival rate supported by the system under a set of constraints), while not preventing the GO user from achieving its target goal performance. This opens new ways of sharing spectrum resources and defining achievable rate regions of interference channels, towards a goal-oriented and compute resource-aware use of the spectrum, as initially suggested in [4]. During a slot t of duration τ, we approximate UE_d average rate as [6]:

$$R_d(t) = \frac{1}{\tau}\big[D_{\text{tx}}(t) W \log_2(1+\text{SINR}_d(t)) + (\tau - D_{\text{tx}}(t)) W \log_2(1+\text{SNR}_d(t))\big], \quad (3)$$

where the first term accounts for the GO transmission period and the second term for the remaining portion of time[2], with $\text{SNR}_d(t)$ the signal-to-noise ratio obtained by removing the interference term in (2). Then, we assume that UE_d generates a continuous flow of data, with new arrivals $A_d(t)$ (in bits) at time t being stored in a buffer before transmission. As clarified in the sequel, inspired by [7]–[9], the goal is to maximize these arrivals while guaranteeing queue stability, considering UE_d as equipped with an infinite-size buffer that evolves as

$$Q_d(t+1) = \max(0, Q_d(t) - \tau R_d(t)) + A_d(t), \quad (4)$$

where τ denotes the slot duration, and $A_d(t)$ the number of arrivals *admitted* to the queue during time slot t. We define the sustained data arrival rate as $\bar{A}_d$ (cf. (1)), from which, under the assumption of strong stability (to be guaranteed via the optimization in Sec. III), by Little's law as in [9], we can compute the average queuing delay as $\bar{D}_{q,d} = \tau \bar{Q}_d / \bar{A}_d$.

B. Goal-effectiveness and proactive batch drop

Since both users are served by the same AP, we assume the latter to orchestrate resources. Thus, the modulation scheme employed by UE_g and the wireless channels by the AP, which needs to select an AI model for inference, allocate computing resources for GO user data inference, and allow the DO user to communicate on the same spectrum with a selected transmit power. The latter affects the quality of the received data from the GO user, and, consequently, the performance of the inference task. Since the transmission delay is known thanks to the knowledge of the modulation order, we assume the overall delay for UE_g to be known by the AP (or, accurately predictable) at time t. Then, we assume that a data batch can be proactively dropped if it cannot be treated within a predefined deadline $D_{\max}$, or simply to allow the DO user to improve its performance. This proactive dropping policy depends on the *altruism* of the UE_g to sacrifice inference quality, with the objective of enhancing UE_d performance. From a protocol perspective, the AP takes this decision based on a negotiated service level agreement with the GO user, with the altruism depending on several factors including, among others, environmental concerns by the GO user, or even more convenient monetary costs agreed with the operator. To model the drop decision, we denote by $\gamma(t) \in \{0,1\}$ a variable that equals 0 if the batch is dropped at time t. Then, we define the *instantaneous* goal-achievement as $\Gamma_g(t) = \Gamma_{l(t)}(P_b(t)) \cdot \gamma(t)$, with the goal-effectiveness being $\bar{\Gamma}_g$ (cf. (1)).

*Definition 1 (**Goal-effective achievable rate region**):* Given network conditions (e.g., wireless channels, computing resources, data arrivals), resource allocation, and constraints, it is the set of pairs $(\bar{A}_d, \bar{\Gamma}_g)$ achievable by the system.

We now propose a problem formulation and solution able to efficiently explore the GEARRs and their boundaries.

III. Problem formulation & solution

Maximizing the average data rate of UE_d is a challenging task, due to the variability of data arrivals, wireless channels, and long-term constraints on goal-effectiveness and compute resource usage. We propose a similar approach as proposed in [9], in which queuing theory and stochastic optimization are exploited to maximize the average throughput of a multi-user network under long-term constraints. First, as introduced in Sec. II and illustrated in Fig. 1, UE_d is equipped with a buffer to store bits before transmission (cf. (4)). The first requirement is for this queue to be strongly stable, i.e., $\bar{Q}_d < \infty$ (cf. (1)). Strong stability is achieved if the departure rate (i.e., $\bar{R}_d$) is greater than the arrival rate (i.e., $\bar{A}_d/\tau$). This condition can be achieved by either increasing the departure rate (i.e., the average UE_d data rate), or decreasing the arrival rate. The former can be increased by increasing DO user transmit power, and thus interference to the GO system, while the latter can be achieved via an engineered *proactive* packet/bit drop policy. This results in a fictitious *control valve* (cf. Fig. 1) that chokes arrivals, thus matching the arrival rate to the *goal-effective capacity* of the system. Therefore, our objective translates into maximizing the arrival rate of UE_d, under long-term constraints on: *i)* its buffer stability, *ii)* a goal-effectiveness threshold for UE_g, *iii)* an average constraint on the number of FLOPS performed by the ES. In each slot, the optimization variables are: *i)* the data arrivals admitted to the DO user buffer, *ii)* the UE_d transmit power $p_d^{\text{tx}}(t)$, iii) the GO user proactive batch drop, and iv) the allocated computing resources (FLOPS). The long-term problem is formulated as follows:

$$\max_{\{\varphi(t)\}_{\forall t}} \ \bar{A}_d \quad (5)$$

$$\text{subject to} \quad \textbf{(a)}\ \bar{Q}_d < \infty, \quad \textbf{(b)}\ \bar{\Gamma}_g \geq \bar{\Gamma}_{g,\text{th}}, \quad \textbf{(c)}\ \bar{F} \leq \bar{F}_{\text{th}},$$
$$\textbf{(d)}\ 0 \leq A_d(t) \leq A_d^{\max}(t),\ \forall t, \quad \textbf{(e)}\ p_{\text{tx},d}(t) \in \mathcal{P},\ \forall t,$$
$$\textbf{(f)}\ \gamma \in \{0,1\},\ \forall t, \quad \textbf{(g)}\ \gamma(t) D_{\text{tot}}(t) \leq D_{\max},\ \forall t,$$
$$\textbf{(h)}\ l(t) \in \mathcal{L},\ \forall t, \quad \textbf{(i)}\ 0 \leq F(t) \leq F_{\max},\ \forall t.$$

Besides long-term constraints **(a)**-**(c)** on queue stability, goal-effectiveness and average compute resource load, the instantaneous constraints have the following meaning: **(d)** the admitted arrivals to the queue are non negative and below the actual arrivals at time t; **(e)** the UE_d transmit power belongs to a predefined discrete set $\mathcal{P}$; **(f)** a batch is either dropped or

[2]We assume that one inference data point (or batch) is uploaded by UE_g per time slot. More involved inference traffic profiles are left for future work.

transmitted; **(g)** if transmitted, a batch is treated within the delay threshold; **(h)** the selected inference model belongs to the set of available models; **(i)** $F(t)$ the allocated compute resources are non negative and below a maximum value.

Problem (5) is challenging due to its long-term nature (in terms of objective function and constraints), especially in the absence of a priori statistical knowledge, and non-convexity.

Proposed solution. We propose to solve the problem by transforming (5) into a pure stability problem [7]. The latter concerns the buffer, and two *virtual queues* for constraints (b)-(c), whose time evolution is respectively defined as follows:

$$Z(t+1) = \max\left(0, Z(t) - \mu_z\left(\Gamma_g(t) - \bar{\Gamma}_{g,\text{th}}\right)\right) \quad (6)$$

$$Y(t+1) = \max\left(0, Y(t) + \mu_y\left(F(t) - \bar{F}_{\text{th}}\right)\right), \quad (7)$$

with $\mu_{z(y)} > 0$. Following [7]–[9], virtual queues mean rate stability[3] is a sufficient condition for guaranteeing the associated constraint, through the definition of the Lyapunov function $L(\mathbf{q}(\mathbf{t})) = \frac{1}{2}Q_d^2(t) + \frac{1}{2}Z^2(t) + \frac{1}{2}Y^2(t)$, with $\mathbf{q}(\mathbf{t})$ denoting a vector that contains all queues (physical and virtual). The mean rate stability is guaranteed by a bounded *drift-plus-penalty* function, which is defined as follows:

$$\delta_p(t) = \mathbb{E}\left\{L(\mathbf{q}(t+1)) - L(\mathbf{q}(t)) - VA_d(t)|\mathbf{q}(t)\right\}, \quad (8)$$

with V a trade-off parameter used to balance queue stability (i.e., DO user delay and constraint violations) and objective function (in this case data arrivals, i.e., DO user data rate). The higher the value of V is, the closer to optimal the solution is, with a cost on queueing delay for the DO user. As in [9], we proceed by instantaneously minimizing an upper bound of (8), only based on current observation of wireless channels, GO user modulation scheme, and data arrivals (we omit the derivations due to the lack of space). The problem can be split.

1) First sub-problem - optimal data arrivals control

$$\max_{0\leq A_d(t)\leq A_d^{\max}(t)} (V - Q_d(t))A_d(t) \quad (9)$$

(9) is a linear problem that can solved in closed form, and its optimal solution is $A_d^*(t) = A_d^{\max}(t) \cdot \mathbf{1}\{Q_d(t) \leq V\}$.

2) Second sub-problem - $p_{tx,d}(t), \gamma(t), l(t), F(t)$

The second sub-problem is solved to select $p_{\text{tx},d}(t)$, the inference model $l(t)$, and $F(t)$. It is formulated as follows:

$$\min_{\{p_{\text{tx},d},\gamma,l,F\}} -Q_d(t)\tau R_d(t) - \mu_z Z(t)\cdot\Gamma_g(t) + \mu_y Y(t)\cdot F(t)$$
$$\text{subject to } \textbf{(e)}\text{-}\textbf{(i)} \text{ of (5)} \quad (10)$$

Problem (10) is a mixed-integer non-linear program, however extremely simplified with respect to (5), as the long-term horizon disappears, and only instantaneous search is needed. Then, assuming a limited number of inference models, we can perform an exhaustive search over $l(t)$, and $p_{\text{tx},d}(t)$, and $\gamma(t)$ to subsequently select $F(t)$. We note that, while it is expected the number of models not to exceed a few units (e.g., due to memory constraints), the presence of more users could make this exhaustive search not scalable. Future investigations will deal with this potential issue via solutions based on multi-agent deep reinforcement learning. Concerning the function $\Gamma_l(P_b)$ linking reliability to inference model and the BER, we estimate it a priori on a validation set. Numerical results will clarify the effectiveness of this procedure, with evaluation on a different test set. Once the model and the transmit power are set, the optimal computing power $F(t)$ is computed as the minimum value guaranteeing constraint (g) of (5). If the latter cannot be met, the batch is dropped as no timely inference can be performed. This search over the limited set is possible thanks to the decoupling of the problem over time. Then, once resources are optimized, communication occurs for the two users, computation takes place for the GO user, and all queues are updated (cf. (4), (6)). Finally, the next slot is visited.

[3]For a virtual queue $Z(t)$, it is defined as $\lim_{T\to\infty}\mathbb{E}\{Z(T)\}/T = 0$.

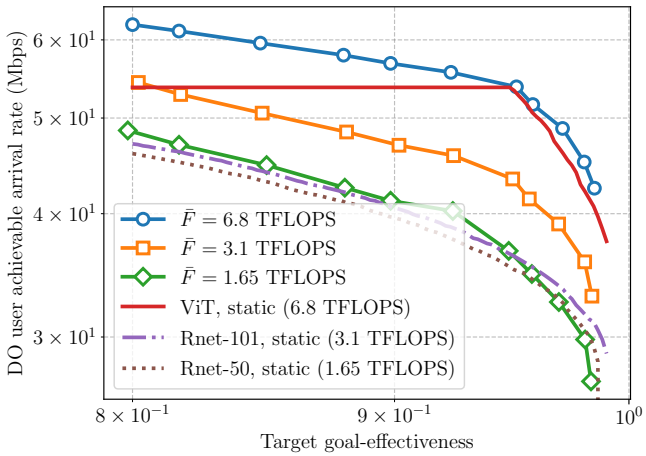


Fig. 2: Goal-effective achievable rate regions

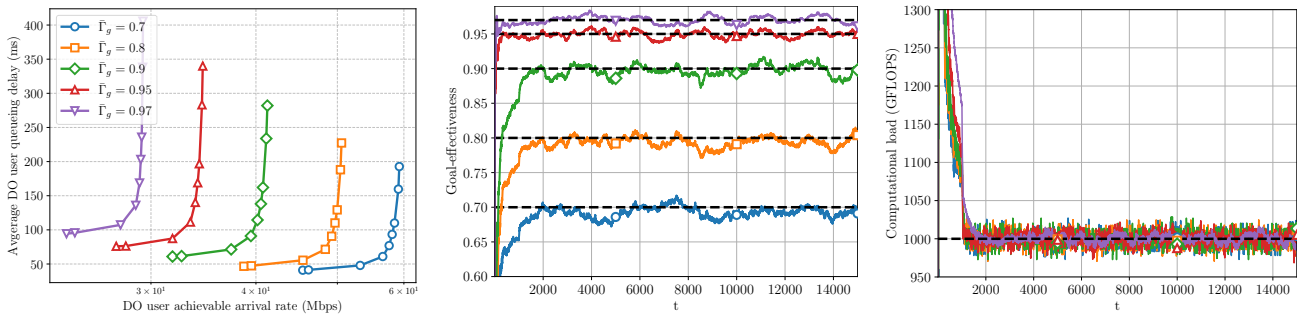
IV. Numerical Evaluation

TABLE I: Simulation parameters

Parameter	Value
carrier freq. (GHz)/W (MHz)/$\mathcal{M}$	3.5/10/256-QAM
noise PSD (dBm/Hz)/noise figure (dB)	-174/10
channel model/# of AP antennas	Rician ($K = 4$), path loss exp. 3.5/8
DO/GO/AP positions (x, y) [m]	[-15, 0]/[0, 0]/[0, 20]
DO/GO user transmit power	$p_{\text{tx},d} \in [0, 0.1]$ W/$p_{\text{tx},g} = 0.1$ W
$D_{\max}$ (ms)/τ (ms)/$A_{\max}$ (bits)	20/20/Poisson, $\lambda = 5\times10^6$
$\mathbf{w}_{g(d)}$ (AP combining vector)	Maximum Ratio Combining
Inference models & their comp. load (GFLOPs) Task & Dataset	Mobilenetv3-small, Resnet-50/101, vit_b_16 0.11, 8.2/15.6, 33 Classification on imagenette [10]

We now show our method's capability of achieving the GEARRs boundaries and exploring the desired trade-offs between GO, DO performance, and computational cost. Simulation parameters are reported in Table I. Simulations are run for 20000 slots and averaged over the last 10000.

First, we assess the achieved performance of our method in terms of GEARRs, to show its capability to get similar or even better performance compared to a *static* policy (fixed DO transmit power and inference model), depending on the computational load constraints. For the latter, given a goal-effectiveness constraint, the best data rate is found via an a *posteriori exhaustive search*, and the corresponding computational resources are selected to meet the delay constraint. Then, it should be noted that the *static* policy requires an exhaustive



(a) Trade-off delay - sustained arrival rate (b) Evolution of goal-effectiveness (c) Evolution of computational load

Fig. 3: Trade-off between DO user sustained data rate, goal-effectiveness, and computational load

search over the set $\mathcal{P}$ of DO user transmit power, after statistical parameters are explored. To fairly compare performance, we set constraint (c) in (5) to the values that are needed by the *static* policy to achieve the target goal-effectiveness (computed a posteriori). In Fig. 2, we show the GEARRs achieved with our method under the different computational constraints, against the static policy with Resnet-50/101 and vit_b_16. As we can notice, the method is able to achieve better performance than the static exhaustive search, with the same respective computational load, with larger gain for lower goal-effectiveness targets. This is thanks to the dynamic decisions based on instantaneous parameters, which allow the system to explore more convenient solutions in the long-term sense (e.g., exploiting favourable channel conditions), and to the proactive GO dropping policy. These degrees of freedom shrink when imposing higher goal-effectiveness constraints, thus making the method approach the boundaries of the GEARRs that are obtained via the exhaustive search. However, this is achieved in a dynamic way, *only based on instantaneous observations* and without the need to estimate the statistics of the involved variables, which makes the method more suitable for being deployed and work online. This first example shows the capability of our method to achieve the boundaries of the GEARRs by dynamically selecting cross-layer parameters, and guaranteeing all the required long-term constraints. To further show the flexibility of this framework, in Fig. 3, we show:

(a) the average DO user queueing delay as a function of its achieved arrival rate, under different goal-effectiveness constraints (the different curves), with $\bar{F}_{\text{th}} = 1$ TFLOPS;
(b) the evolution of the goal-effectiveness, averaged over a 10^3 samples moving window, for the different thresholds;
(c) The evolution of the average computation resources (cf. constraint **(i)** of (5)) using a 10^3 samples moving window.

From Fig. 3a, we can notice that, for a given goal effectiveness constraint for the GO user, the queuing delay at the DO user increases as a function of the achieved data arrival rate, as predicted by the theory [7], with an asymptotic behaviour around the maximum value (boundary of the GEARR under the imposed constraint). Obviously, a stricter goal-effectiveness requirement results in degraded DO user performance, showing the complex multi-dimensional trade-off involving different layers. Finally, Figs. 3b and 3c show the convergence of the goal-effectiveness (on the test set using the reliability function estimated on the validation set) and the average computational load toward the desired values.

V. Conclusion

We proposed a novel cross-layer and cross-domain resource allocation framework, under which interference-prone spectrum sharing is managed based on higher layer parameters belonging to the world of edge intelligence, namely the diverse model inference capabilities and computational load awareness. Our findings suggest that computing power and inference model robustness to bit-level errors can help boosting the performance of legacy users that use the spectrum to maximize classical metrics such as the data uploading rate. Based on these findings, we proposed a dynamic method that jointly encompasses communication, computing and edge AI aspects, toward a computation- and goal-aware spectrum sharing.

References

[1] E. C. Strinati *et al.*, "Goal-oriented and semantic communication in 6G AI-native networks: The 6G-GOALS approach," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 1–6.

[2] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 155–169, 2023.

[3] Y. Liu and B. Clerckx, "Rate-splitting multiple access for coexistence of semantic and bit communications," *Online: https://www.arxiv.org/pdf/2409.10314*, 2022.

[4] M. Merluzzi, M. C. Filippou, L. Gomes Baltar, M. D. Mueck, and E. Calvanese Strinati, "6G goal-oriented communications: How to coexist with legacy systems?" *Telecom*, vol. 5, no. 1, pp. 65–97, 2024. [Online]. Available: https://www.mdpi.com/2673-4001/5/1/5

[5] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Press, 2005.

[6] M. C. Filippou, D. Gesbert, and G. A. Ropokis, "A comparative performance analysis of interweave and underlay multi-antenna cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2911–2925, 2015.

[7] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morg. and Clay. Pub., 2010.

[8] S. Lakshminarayana and T. Q. Quek, "Throughput maximization with channel acquisition in energy harvesting systems," in *2014 IEEE Int. Conf. on Communications (ICC)*, 2014, pp. 2430–2435.

[9] M. Merluzzi, S. Bories, and E. C. Strinati, "Energy-efficient dynamic edge computing with electromagnetic field exposure constraints," in *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2022, pp. 202–207.

[10] J. Howard, "Imagenette: A smaller subset of 10 easily classified classes from ImageNet," March 2019. [Online]. Available: https://github.com/fastai/imagenette