

Article

6G Goal-Oriented Communications: How to Coexist with Legacy Systems?

Mattia Merluzzi ^{1,*}, Miltiadis C. Filippou ^{2,†}, Leonardo Gomes Baltar ^{2,‡}, Markus Dominik Mueck ² and Emilio Calvanese Strinati ¹

¹ University Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France; emilio.calvanese-strinati@cea.fr

² Intel Deutschland GmbH, 85579 Munich, Germany; miltiades.filippou.de@ieee.org (M.C.F.); lbaltar@ieee.org (L.G.B.); markus.dominik.mueck@intel.com (M.D.M.)

* Correspondence: mattia.merluzzi@cea.fr

† Current affiliation: Nokia Strategy & Technology, 81541 Munich, Germany.

‡ Current affiliation: Volkswagen Infotainment GmbH, 44803 Bochum, Germany.

Abstract: 6G will connect heterogeneous intelligent agents to make them natively operate complex cooperative tasks. When connecting intelligence, two main research questions arise to identify how artificial intelligence and machine learning models behave depending on (i) their input data quality, affected by errors induced by interference and additive noise during wireless communication; (ii) their contextual effectiveness and resilience to interpret and exploit the meaning behind the data. Both questions are within the realm of *semantic* and *goal-oriented* communications. With this paper, we investigate how to effectively share communication spectrum resources between a legacy communication system (i.e., data-oriented) and a new goal-oriented edge intelligence one. Specifically, we address the scenario of an enhanced Mobile Broadband (eMBB) service, i.e., a user uploading a video stream to a radio access point, interfering with an edge inference system, in which a user uploads images to a Mobile Edge Host that runs a classification task. Our objective is to achieve, through cooperation, the highest eMBB service data rate, subject to a targeted *goal effectiveness* of the edge inference service, namely the probability of confident inference on time. We first formalize a general definition of a goal in the context of wireless communications. This includes the goal effectiveness, (i.e., the goal achievability rate, or the probability of achieving the goal), as well as *goal cost* (i.e., the network resource consumption needed to achieve the goal with target effectiveness). We argue and show, through numerical evaluations, that communication reliability and goal effectiveness are not straightforwardly linked. Then, after a performance evaluation aiming to clarify the difference between communication performance and goal effectiveness, a long-term optimization problem is formulated and solved via Lyapunov stochastic network optimization tools to guarantee the desired target performance. Finally, our numerical results assess the advantages of the proposed optimization and the superiority of the goal-oriented strategy against baseline 5G-compliant legacy approaches, under both stationary and non-stationary communication (and computation) environments.

Keywords: goal-oriented communications; 6G; connecting intelligence; edge inference; edge artificial intelligence



Citation: Merluzzi, M.; Filippou, M.C.; Gomes Baltar, L.; Mueck, M.D.; Calvanese Strinati, E. 6G Goal-Oriented Communications: How to Coexist with Legacy Systems? *Telecom* **2024**, *5*, 65–97. <https://doi.org/10.3390/telecom5010005>

Academic Editors:
Alexandros-Apostolos Boulogeorgos,
Panagiotis Sarigiannidis, Thomas
Lagkas, Vasileios Argyriou and
Pantelis Angelidis

Received: 6 October 2023
Revised: 24 November 2023
Accepted: 5 January 2024
Published: 24 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Today, as the fifth mobile generation (5G) is at its deployment stage (<https://www.3gpp.org/technologies/5g-system-overview> accessed on 20 July 2023), the race to 6G has started all around the world [1,2]. The latter entails theoretical research of new technologies' fundamental limits, the definition of new cutting edge use cases [3,4], their associated Key Performance Indicators (KPIs) and Key Value Indicators (KVI) and, obviously, the new technological levers to enable the ecosystem. New technological enablers are required at all layers of the protocol stack, from application and network to the physical

layer [5,6], and also down to the wireless propagation environment with reconfigurable intelligent surfaces [7–9]. This cross-layer perspective will give birth to a fully reconfigurable system that can adapt to extremely diverse requirements, with a joint orchestration of heterogeneous resources (wireless, computing, storage). Following the revolution provided by Multi-access Edge Computing (MEC) [10], a paradigm shift in 6G will be the integration of communication, computing and storage, natively enabling *heterogeneous intelligence* (e.g., humans, robots or machines) to accomplish complex cooperative tasks [11–15]. Connecting intelligence requires informative cooperative interactions. New trends in 6G explore both the *semantic problem* of understanding the meaning of the source data and the *effectiveness problem*, whose aim is accomplishing target goals through possibly corrupted/compressed/encoded/misunderstood received information [16].

Achieving goals with target effectiveness requirements defines a new family of KPIs in future 6G systems. We refer to such family of KPIs as *goal effectiveness*. This paper explores as main KPIs the *goal effectiveness* together with its associated *goal cost*, which refers to the price to pay to achieve a goal. Goal effectiveness is one of the main KPIs to assess the performance of *goal-oriented (GO) communication* [16,17], a true paradigm shift from *data-oriented (DO) communication* [11], envisioned to proliferate in 6G. While the aim of DO communications is for an intended data destination to reliably receive the original data generated by a data source, the aim of GO communications is rather for the data destination (playing the role of an acting agent) to receive adequate data of sufficient quality in the context of correctly performing a (possibly cooperative) task. Indeed, in our view, GO communications can refer to:

- *goal-oriented data compression* [11,18–28], aiming at extracting and adapting the relevant information needed to make the receiver accomplish a goal with desired effectiveness. This can be based on semantic information extraction [22,28,29], but it is not restricted to the meaning of data;
- *goal-oriented transmission* [30–32], aiming at adapting communication reliability (e.g., the Packet Error Rate—PER), to achieve target goal effectiveness, i.e., the probability of achieving the goal. This can also involve semantic-aware packet protection, under the assumption that some packets bring more relevant information than others, from the perspective of the goal.

Both concepts above also entail the definition of costs, which can include, among the others, energy consumption, delay, resource utilization and communication overhead. Obviously, a combination of both approaches is possible and would represent a unified view towards a fully GO system design.

As a second aspect tackled in this paper, a fundamental challenge of future 6G systems is to accommodate, on the same network infrastructure, services that are heterogeneous from several perspectives: requirements, resource consumption, time-variability, etc. Therefore, while it is true, in our view, that 6G will enable new classes of GO and semantic services, it is also certain that it will keep serving classical users and/or *verticals* [33] (i.e., data-oriented services), which, in the sequel, we refer to as *legacy communication or, data-oriented systems and services* (in this paper, we will refer to them as *DO systems* for brevity). For this reason, at some point in the near future (e.g., in the next 10 years), a natural question will arise on the coexistence of GO and DO systems, to understand how one system affects the other, when they share the same network infrastructure and wireless resources. A possible question in this context is:

How does the interference of a DO system affect GO performance, and at which cost?

A similar question is considered in [34], where the authors analyze the performance of a system in which a semantic and a DO service coexist.

In this paper, we aim to answer the question of coexistence of GO and DO systems, based on a formal definition of a goal, entailing effectiveness and costs that jointly take into account GO and DO system requirements. After a general yet formal definition of goal and GO optimization problems, we will consider the case of MEC-enabled edge inference as a use case for the GO system. In the proposed setting, a GO system aims to classify data

collected by an end device and transmitted to a nearby Mobile Edge Host (MEH), while a DO system aims to upload data with the highest possible data rate. The fundamental question to answer is then the following:

What is the highest DO system data rate that can be achieved while, at the same time, guaranteeing a target goal effectiveness of the GO system, assuming full spectrum sharing (i.e., in the presence of co-channel interference)?

We believe that this and other related questions will arise in the future for several use cases. Among the others, due to the bonding between communication and computing, we can mention the edge intelligence paradigm for robot remote control in factory plans, environment (e.g., factory floor) digital twins for manufacturing, identification of faulty parts in a production line, vehicular communications for cooperative perception and the coexistence of ground and aerial communications [4]. For the latter, the coexistence pertains not only to GO and DO services but also to terrestrial and non-terrestrial communications, which makes system optimization even more challenging. Of course, for such new services and use cases, different metrics beyond the data rate can be taken into account and traded off with the goal effectiveness of coexisting GO services.

1.1. Related Work

Although the semantic and GO communication paradigm is in its infancy within the communication community, several works already started investigating the fundamental limits and the algorithmic foundations. In [35], the authors propose an end-to-end semantic communication model, entailing sampling, filtering, preprocessing, reception and control. In [18], a GO data quantization scheme is proposed to reduce the communication overhead of a decision-making service. Similarly, a GO data quantization and data clustering approach is proposed in [20]. In [19], the authors propose an online algorithm able to adapt to channel conditions through the selection of a GO compression scheme in an edge inference scenario. Similarly, reference [21] proposes a GO data compression method based on the information bottleneck principle. The authors of [22] show the gain of a semantic-aware transmission on the reconstruction error of a real-time tracking system, while [23] focuses on the cooperation between agents with a common goal through a task-oriented mutual information exchange. In [24], it is shown how cooperation of different Machine Learning (ML) models can help improve the energy efficiency of an edge inference system with quantized data transmission, and a similar GO resource allocation based on an adaptive selection of JPEG compression is proposed in [25]. In [27], a goal-aware DNN splitting and feature extraction is proposed in an edge inference scenario, while in [28], a unified GO semantic communication framework is proposed in the context of a visual question answering use case. From a GO transmission perspective, the authors of [30] propose a joint transmission and recognition scheme, showing the effect of communication reliability on the accuracy, however not considering delay and energy consumption, as well as cooperative inference. Moreover, in our precursor conference paper [31], we investigate the performance of communication KPIs (i.e., the bit error rate) on a real-time edge inference task, taking into account classification accuracy, energy and delay. Furthermore, reference [34] introduces, for the first time, the concept of coexistence between semantic and DO services, with a performance evaluation aimed to achieve sum-rate maximization of DO users, with a minimum required SNR constraint for semantic users, with mutual interference. However, the authors of [34] do not consider goal-effectiveness constraints in the presence of interfering DO users and do not propose a GO optimization, which is the purpose of this paper for an edge inference use case.

Finally, references [32,36], which are the most closely related works, exploit semantic communication for Non-Orthogonal Multiple Access (NOMA) by considering a system with a primary and a secondary user served by the same Access Point (AP) through NOMA. The authors show the gain of defining a semantic ergodic rate in the overall system performance, with the aim of maximizing the primary (i.e., bit-level communication) user data rate, under a semantic ergodic rate constraint of the secondary user, controlling

transmit power and transmission time scheduling. Starting from the research question of [32,36], we investigate a novel scenario, in which a GO and a DO communication system coexist, thus focusing on GO, rather than semantic, communication.

1.2. Our Contribution

Similarly to [32,36], in this paper, we aim at analyzing the impact of mutual interference on future service classes. Differently from [32,36], whose focus is on semantic communication and ergodic (semantic) rate, and where a single base station serves users with bit and semantic streams, we rather focus on GO communication and resource allocation in connect-compute networks, in which a GO and a DO communication system coexist. Therefore, the main contributions of our work are the following:

- We consider the communication effectiveness in accomplishing a predefined goal/task as our system constraint. Therefore, the focus of our work is on GO communications (i.e., effectiveness of communication toward achieving a goal—and in particular through resource allocation) rather than semantic communications (i.e., understanding the meaning of data).
- We introduce the edge inference service as the main use case under investigation, defining it as a GO communication service and problem.
- We consider the communication reliability (i.e., Packet Error Rate—PER) as a variable to be controlled to achieve target goal effectiveness at the GO system while maximizing performance of the DO system in terms of data rate.
- We propose a *computation resource-aware* method for guaranteeing goal effectiveness, taking into account the computing resource availability for edge inference.

In other words, our work is about goal-aware system coexistence, while [32,36] are about a multiple access scheme where different traffic types (semantic, bit-oriented) are multiplexed by a single AP that can process both sorts of data.

Then, as a first step of this work, we formally define a GO resource optimization problem in the context of wireless communication systems. This will help us to introduce, in the sequel, the fundamental problem of coexistence of GO and DO systems in 6G, with full spectrum sharing. As a specific use case, we consider an edge inference service for the GO system, in which an end device uploads data to a MEH, which runs a classification task through a pre-trained and pre-uploaded ML model (here, a Convolutional Neural Network—CNN). In such a system, we will define goal effectiveness and goal cost, with the former relating to the probability of receiving confident classification results within a predefined deadline and the latter relating to the data rate loss of the DO user, compared to the case in which the GO user is absent. This bonds interference, communication reliability, computing resources, goal effectiveness and goal cost into a unified framework. To the best of our knowledge, this has never been studied before in the literature for GO communication and edge inference scenarios jointly factoring in wireless and compute resources.

After a performance evaluation of the proposed scenario, aiming to gain insights about the system behavior as a function of different free system parameters, we will propose a simple yet relevant optimization problem, along with an adaptive algorithm able to optimize radio resources to minimize the goal cost, under goal-effectiveness constraints, with a connect-compute resources-aware approach. In summary, the main novelty of the paper consists of conducting a GO performance analysis of an edge inference system sharing resources with a DO system, along with an adaptive algorithm to attain desired performance, being aware of application constraints, its online performance, as well as connect-compute resource availability. After proposing an algorithm with theoretical guarantees, we test it on scenarios with different parameters and requirements, against baseline strategies (e.g., orthogonal bandwidth splitting), but also on non-stationary environments in which GO requirements (e.g., goal effectiveness) or computation resource availability statistics can unexpectedly change over time.

1.3. Organization of the Paper

The remainder of this paper is organized as follows: Section 2 provides a general definition of a goal in the context of wireless communications, entailing effectiveness, constraints, costs and goal achievability; Section 3 focuses on applying the definition to the proposed scenario, in which GO and DO systems coexist, fitting the general definitions to the specific use case of edge inference (GO user) and data upload (e.g., video streaming—DO system). In addition, a numerical evaluation without optimization is provided to explore performance and gain insights about the relevant parameters that will be considered as optimization variables in Section 4, in which a long-term problem formulation, along with a solution based on Lyapunov stochastic network optimization, is proposed. Numerical results assessing the performance of the proposed algorithm are presented in Section 5, while Section 6 draws the conclusions of the paper and proposes some future research directions.

1.4. Notation and Acronyms

Throughout the paper, bold lower case and upper case letters denote vectors and matrices, respectively; the operator $|\cdot|$ is the absolute value of a complex number, and the superscript H denotes the Hermitian operator; $card(\cdot)$ denotes the cardinality of a set, and $\mathbf{1}_A$ denotes the indicator function, which equals 1 if condition A is satisfied and 0 otherwise. Finally, calligraphic upper case letters always denote the long-term average of a corresponding variable throughout the text; for instance, for variable x (or X), the long-term average is denoted as \mathcal{X} , and it reads as follows:

$$\mathcal{X} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{x_t\}, \quad (1)$$

where t indicates a temporal index. The acronyms used throughout the paper are defined the first time they appear but can be also found in Table 1, in alphabetic order.

Table 1. List of acronyms.

AI	Artificial Intelligence	GPU	Graphical Processing Unit
AP	Access Point	KPI	Key Performance Indicator
CLD	Conditional Lyapunov Drift	KVI	Key Value Indicator
CNN	Convolutional Neural Network	MEC	Mutli-access Edge Computing
CPU	Central Processing Unit	MEH	Mobile Edge Host
DNN	Deep Neural Network	MCS	Modulation and Coding Scheme
DO	Data-Oriented	ML	Machine Learning
DPP	Drift Plus Penalty	NREI	Negative Relative average Entropy Increase
DRL	Deep Reinforcement Learning	NOMA	Non-Orthogonal Multiple Access
E2E	End-to-End	PER	Packet Error Rate
eMBB	enhanced Mobile Broad Band	SINR	Signal-to-Interference-plus-Noise Ratio
GO	Goal-Oriented	UE	User Equipment

2. Definition of a Goal

The GO communication paradigm constitutes a communication approach for which performance is not measured by classical metrics, such as data rate or wireless link reliability, but rather by the success level with which a network entity is able to perform a sequence of application-related tasks, as a result of exchanging data with one or more other network entities. In the sequel, we will also refer to these entities as *agents*. In this context, an agent is any entity endowed with communication capabilities, which could also have computing capabilities and be possibly embarked with Artificial Intelligence (AI) and/or ML models.

In the latter case, we can refer to it as an *AI agent*. However, this distinction is not essential in this paper.

A fundamental step is to formally define a goal to be able to properly formulate GO communication problems. Indeed, a wrong, inexact, unclear or ambiguous definition of a goal may lead to dramatically wrong or biased decisions and communication/computation policies.

In [37], E. Hollnagel defines a goal out of the context of wireless communications. From its definitions, and with the aim of embedding the GO perspective into wireless communication systems, in the most general case, a goal is characterized by a set of requirements that, if guaranteed, determine its accomplishment. When embedded in wireless scenarios with inter-agent communications, the accomplishment of these requirements is directly linked to communication (and computation)-related performance and strategies. As an example, suppose that the goal of the communication is to exchange sampled data for anomaly detection. Then, data corrupted by a noisy (and possible interference-prone) channel could lead to wrong decisions, depending on the wireless link reliability (e.g., the PER). At the same time, data encoding schemes (e.g., compression) affect the goal achievement due to distortion. Both data encoding and wireless link reliability play a key role in GO communications. A priori, it is not generally easy to predict the exact correlation between goal accomplishment and wireless communication performance, data encoding, etc. This depends on several aspects, including the goal definition, the connect-compute network conditions and the a priori knowledge of communicating agents. Learning and timely adapting communication parameters towards goal accomplishment is the main target of the GO communication paradigm. As also remarked in [37], a goal is achieved through a series of tasks, with a task being a collection of actions. Generally, an action entails a decision on a set of parameters, which could involve communication, computation and possibly control. Let us formally define the three measures that characterize a goal and are necessary to formulate GO communication problems.

2.1. The Goal Value, Achievability and Effectiveness

As anticipated, the first fundamental step is for an agent to be able to verify whether a goal has been achieved by carefully selecting metrics assessed by the agent for that purpose. In this context, we define a quantified representation of the target system state indicating goal accomplishment, which we name as *goal value*. The goal value can be, for instance, a binary variable that equals 1 if the goal is achieved and 0 otherwise. Of course, it is not restricted to this case, and it may be represented as a generic real vector (e.g., the position in space of a robot with respect to a target one). In a dynamic system, a goal value is possibly time-varying, i.e., at each time instant t , a new/updated goal value can be made available to an agent as a result of environment states, agents' actions and environment feedback. Therefore, in the most general case, the goal value can be denoted as a real variable $\Theta_t \in \mathbb{R}^l$, with t denoting the time instant.

Goal Achievability and Goal Effectiveness

Going beyond the goal value, and, with focus on either repetitive or long-standing application-related goals, agents eventually have to measure the goal accomplishment rate, i.e., the probability of achieving the goal value. Indeed, a goal accomplishment must not necessarily be deterministic. For instance, a goal value can be obtained with a certain probability, which we can refer to as goal effectiveness. The goal effectiveness is inherently non-negative and is a long-term measure. Indeed, even in the case that a goal is a sequence of sub-goals, the sub-goal accomplishment is stochastic, and what we are interested in is the success rate of the goal. We can give the following example: (i) edge pattern (e.g., image) classification: in this use case, we can choose the goal value to be equal to 1 if a pattern is correctly classified and 0 otherwise. In this case, without additional constraints, the goal effectiveness is the probability that the goal value equals 1 and corresponds to the correct classification rate.

Nonetheless, besides the goal value itself, the goal effectiveness may be subject to J instantaneous constraints that, if not met, prevent the goal from being perceived by the agent as accomplished, per a user service-level agreement. Sticking to the edge classification use case, the goal value is defined as before, while a possible constraint is represented by the End-to-End (E2E) delay being kept under a predefined threshold. Without loss of generality, we can write these constraints (which we refer to as *short-term goal constraints*) as $g_{j,t} \leq 0, j = 1, \dots, J$. Building on the goal value and the short-term goal constraints, we can formally define the goal effectiveness as follows:

$$\mathcal{E}_g = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\{ \mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \prod_{j=1}^J \mathbf{1}_{\{g_{j,t} \leq 0\}} \right\}, \quad (2)$$

where Θ_{th} is a predefined goal value threshold, and the expectation is generally taken with respect to the random context parameters and the strategy (policy) followed by the agents to achieve the goal. From (2), we can already state that, as per our definition, the goal effectiveness is a probability, thus $\mathcal{E}_g \in [0, 1]$.

Going further, besides the goal value, achieving a goal may also entail other M long-term constraints. As an example, a typical long-term constraint is to keep the average energy consumption of an agent below a predefined threshold, e.g., to not quickly drain its battery level. In the most general case, we can write these long-term constraints as follows:

$$\mathcal{G}_s := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{G_{s,t}\} \leq \mathcal{G}_{s,th}, \quad \forall s = 1, \dots, S. \quad (3)$$

Building on (2) and (3), we can write the following definition for goal achievability: (i) a goal is *achievable* if there exists at least one policy that guarantees (2) to be above a predefined threshold $\mathcal{E}_{g,th}$; (ii) a goal is *strongly achievable* if there exists at least one policy that guarantees (2) to be above a predefined threshold $\mathcal{E}_{g,th}$, while guaranteeing all other long-term constraints (3). In typical communication system management, long-term constraints entail communication KPIs, e.g., PER below a threshold. This translates into, e.g., adaptive Modulation and Coding Scheme (MCS) selection, able to guarantee the target PER. Instead, with the GO paradigm, *communication* KPIs are not necessarily required to be satisfied a priori, while measures belonging to the physical/human world, or, in general, to the application, are used to assess the performance of the communication system. In this case, classical communication KPIs can be *learnt from experience*, and *dynamically adapted*, to achieve the target goal effectiveness in (2) in the most efficient way, possibly subject to constraints (3).

2.2. The Goal Cost

Every goal should be accompanied by a cost spent to achieve it, i.e., a price to pay in terms of, e.g., radio and computation resources. We refer to this price as *goal cost*. Indeed, there are possibly several strategies achieving the goal with target goal effectiveness, i.e., a goal is possibly achievable via multiple strategies. However, different strategies have also different costs. The aim of GO communications is to find the strategy that achieves the target goal effectiveness with the lowest goal cost. We can differentiate between short-term (or even instantaneous) and long-term costs. In the case of the definition of a long-term cost, we can define an average or a cumulative cost, with the latter being the sum of all instantaneous costs spent to achieve the goal. Normally, the latter refers to a strategy to achieve a goal, where goal success is indicated by reaching a specific terminal state for the (physical) system, e.g., for a robot going from point A to point B, i.e., an *episodic task*, as defined in (Deep) Reinforcement Learning (DRL). On the other hand, the average cost can also be used for those goals that last virtually infinite time, e.g., continuous collection and classification of patterns. Denoting by C_t the instantaneous cost, we can define a long-term cost as follows (cf. (1)):

$$C = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{C_t\}. \quad (4)$$

In what follows, we aim to shed lights on how communication service requirements may impact goal effectiveness and propose formulating the overall GO communication problem involving selection of policies across communication, computation and control domains.

2.3. Identifying Goal-Achieving Communication KPIs

In edge AI/ML scenarios and use cases, multiple agents (at least two), equipped with more or less powerful computing units and complex (ML) models, communicate to achieve a (possibly common) goal. The issue is then how to allocate connect-compute resources in order to guarantee a target goal effectiveness subject to short and long-term constraints, paying the lowest possible goal cost. First, when exchanging data, availability of communication resources affects the level of distortion on the received data with respect to the original information. In classical communication systems, communication KPIs are defined a priori, depending on the different service requirements. This paradigm has started a long race towards adding “9s” to the communication reliability [38] in terms of, e.g., packet success rate. Here, based on the definitions provided in the previous section, we argue that this is not the most efficient way of designing the system and orchestrating operations and resources. Indeed, we do not formulate a problem in which communication KPIs are explicitly taken into account, but they are rather controlled by learning and adaptation to communication policies that achieve target levels of goal effectiveness. In this context, let us denote by π_{comm} a *communication policy* entailing, in the most general case: (i) source encoding (e.g., data compression schemes) [25,26]; (ii) modulation and channel coding [39]; (iii) wireless channel scheduling, including node participation selection [40] and association [41]; (iv) multiple antenna transmission scheme, devising power allocation and precoding [42]. Moreover, let us denote by π_{comp} a *computation policy* entailing, in the most general case: (i) local computing resource scheduling at each device/agent; (ii) computation resource scheduling at shared computing units (e.g., in MEHs) [43–45]; (iii) selection of a single one or multiple collaborative AI agents containing relevant ML models for ensemble-based inferencing [24,46] or federated learning [40]. Finally, we denote by π_{contr} a *control policy* entailing AI agents’ mobility and trajectories.

A general GO communication problem can be then formulated as follows:

$$\begin{aligned} \min_{\{\pi_{\text{comm}}, \pi_{\text{comp}}, \pi_{\text{contr}}\}} \quad & C := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{C_t\} \\ & \text{subject to} \\ & (a) \pi_{\text{comm}} \in \mathcal{P}_{\text{comm}} \\ & (b) \pi_{\text{comp}} \in \mathcal{P}_{\text{comp}} \\ & (c) \pi_{\text{contr}} \in \mathcal{P}_{\text{contr}} \\ & (d) \mathcal{E}_g \geq \mathcal{E}_{g,\text{th}} \\ & (e) \mathcal{G}_s \leq \mathcal{G}_{g,\text{th}}, \quad s = 1, \dots, S, \end{aligned} \quad (5)$$

where $\mathcal{P}_{\text{comm}}$, $\mathcal{P}_{\text{comp}}$, $\mathcal{P}_{\text{contr}}$ denote the feasible sets of communication, computation and control actions, respectively. The constraints in (5) have the following meaning: (a) the communication policy belongs to the feasible set $\mathcal{P}_{\text{comm}}$ (e.g., the transmit power is non-negative and lower than a threshold); (b) the computation policy belongs to the feasible set $\mathcal{P}_{\text{comp}}$ (e.g., the sum of all CPU resources of an MEH assigned to a set of users does not exceed the maximum CPU computing power); (c) the control policy belongs to the feasible set $\mathcal{P}_{\text{contr}}$ (e.g., a robot can move towards a predefined set of possible directions); (d) the goal effectiveness is above a predefined threshold (e.g., the inference accuracy of an edge inference system is above a threshold); (e) other long-term constraints are satisfied (e.g., the average energy of an agent does not exceed a predefined threshold). The problem

formulation in (5) is quite general, and it can be easily customized to specific use cases. In the next section, we present a specific scenario in which a GO and a DO communication system coexist. The aim is to hinge on the GO paradigm and formulation to maximize the DO system data rate under goal-effectiveness constraints of the GO system.

3. Coexistence of a Goal-Oriented and a Data-Oriented Communication System

In this section, we apply the GO approach to a scenario in which two different systems coexist: (i) a GO communication system, consisting of a User Equipment (UE) offloading inference tasks to an MEH collocated to an AP, sharing radio spectrum with (ii) a classical DO communication system, consisting of another UE connected to another AP, where the device continuously uploads content, such as a video stream, to be cached in the network. As an example, the focused scenario can refer to an industrial setting for low-power networking, where a device uploading product images for anomaly detection (GO transmitter/ receiver pair) coexists with a security camera providing continuous video feed of the factory floor (legacy DO eMBB service). We believe that such a scenario will be of extreme interest in the future, as in beyond 5G and 6G systems, and the same system architecture will need to support both such services. As spectrum resources under 6 GHz will continue to be pivotal in providing wide radio coverage [3], (non-orthogonal) spectrum sharing is envisioned to be a standing feature, especially for low-power networking, e.g., in an industrial environment.

The GO approach may find direct application to system scenarios involving coexistence of services of different types. For such services, the most relevant and important communication KPIs (e.g., data rate, latency, reliability) may even be the same, however, the tasks performed may substantially differ in scope, hence design trade-offs need to be addressed to concurrently guarantee all application-related goal values by keeping goal costs as low as possible for all tasks. To further elaborate on the problem of *conflicting goals*, in this paper, we choose to investigate the coexistence of the two systems, inspired by the Cognitive Radio (CR) networking paradigm [47,48], whose scope is to opportunistically use a portion of the spectrum that is licensed to another usage [49]. Recently, the CR paradigm has been investigated in several contexts, including that of relay networks [49,50].

The fundamental difference lies in that, for the CR problem, where a primary and a secondary system operate together in space and time, focus is, in general, on allocating the available secondary radio resources to maximize secondary system performance, subject to the non-violation of *radio* performance constraints imposed by the primary system operator. Nonetheless, for the focused GO/DO system coexistence, the ultimate design target is *service-effective communication*, even with “softened” radio communication requirements for the GO system that can be learned during goal accomplishment.

3.1. System Setup

The system of interest is illustrated in Figure 1. A GO transmitter-receiver pair (which we, in short, term after as GO system), consisting of a single-antenna UE, denoted as UE_g and referred to as *GO user*, and an AP, denoted as AP_g and referred to as *GO AP*, equipped with M_g antennas, fully shares the available radio frequency spectrum of bandwidth W (Hz) with a DO transmitter-receiver pair (which we, in short, term after as DO system) of the same characteristics, i.e., a single-antenna UE, denoted as UE_d and referred to as *DO user* and an AP, denoted as AP_d and referred to as *DO AP*, equipped with M_d antenna elements. It is assumed that AP_g and AP_d are interconnected via a backhaul link characterized by high capacity and low delay. In addition to radio infrastructure, it is assumed that an MEH is collocated with AP_g , where the interconnection delay can be considered as negligible. Generalizing this would be straightforward and is left for future investigations.

confidence, which translates into inference accuracy. These two measures, when jointly considered, define the goal effectiveness for edge inference, as it will be formalized later on in this section.

3.2.1. Uplink Radio Performance of GO User

Let us recall that the GO user offloads inference tasks to the MEH. From a wireless perspective, performance is affected by multiple factors: (i) the uplink data rate; (ii) the interference received from the DO user; (iii) the target packet error rate. Let us denote by $\mathbf{w}_g \in \mathbb{C}^{M_g \times 1}$ the combining vector at AP_g , denoting by P_g the transmit power of UE_g , and by P_d the transmit power of UE_d . Assuming time as organized in slots $t = 1, 2, 3, \dots$, the Signal-to-Interference-plus-Noise Ratio (SINR) at the GO receiver (AP_g), at time t , reads as

$$\text{SINR}_{g,t} = \frac{|\mathbf{w}_{g,t}^H \mathbf{h}_{gg,t}|^2 P_{g,t}}{N_0 W + |\mathbf{w}_{g,t}^H \mathbf{h}_{gd,t}|^2 P_{d,t}}, \quad (7)$$

where index t denotes the time dependence of the involved random and controlled variables; whereas N_0 denotes the power spectral density, and W the uplink bandwidth, assumed to be fully shared between the two systems. Now, considering finite blocklength transmissions [51–53], and denoting by γ_t the target PER at time t , we can write the achievable rate of UE_g (in bits/s) at time t as follows [51]:

$$R_{g,t} = W \left[\log_2(1 + \text{SINR}_{g,t}) - \frac{1}{\log(2)} \sqrt{\frac{V_t}{n_g}} Q^{-1}(\gamma_t) \right], \quad (8)$$

where n_g is the blocklength, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q function, while V_t is the channel dispersion, given by [52,53]

$$V_t = 1 - \frac{1}{(1 + \text{SINR}_{g,t})^2}, \quad (9)$$

which is well approximated as $V \approx 1$ with SINR above 5 dB [52]. Nevertheless, in this work, we keep it general, as our system may be required to work at low SINR regimes, provided that a target goal effectiveness is guaranteed. Finally, assuming (without loss of generality) that a batch of N_t new patterns is requested to be inferred at time slot t , the uplink transmission delay for GO communication reads as follows:

$$D_{\text{tx},t} = \frac{N_t n_b}{R_{g,t}}, \quad (10)$$

where n_b denotes the number of bits encoding one pattern. In this work, we do not focus on the specific encoding scheme, which we assume to be fixed. Semantic and GO compression can also play a role in optimizing n_b , which goes beyond the scope of this paper and is left for future investigations. Preliminary works on online compression level selection are available in [11,24–27,30].

3.2.2. Computation Delay of GO User

Uplink transmission is just the first of the (at least two) phases of wireless edge inference, whose second step is remote processing through an ML model (e.g., a DNN), running in an MEH. The processing phase constitutes another source of delay, which is, typically, non-negligible, especially for demanding inferencing tasks served by MEHs of limited computing capabilities, as compared to the ones at the distant cloud. The computation delay is not, in general, a deterministic quantity, as it depends on CPU loading, execution of background processes, access to memory, etc. Obviously, it strongly depends on the employed ML model [54]. In this paper, we assume that the MEH, irrespective of wireless and computing resource availability, always employs the same model (assumed already trained), and we leave GO model selection criteria for future investigations. As

such, we denote by $D_{\text{comp},t}$ a random variable (whose statistics are possibly unknown) representing the computation time at the MEH. At each time slot, the computing time is possibly different, and its estimation is assumed to be available at the MEH, making it able to optimize connect-compute resources.

Finally, the total inference time, neglecting downlink transmission of small size results (typically a few bits for binary or multi-class classification), is given by

$$D_{\text{tot},t} = D_{\text{tx},t} + D_{\text{comp},t}. \quad (11)$$

Part of the goal is to retrieve inference results earlier than a predefined deadline D_{max} , as clarified in the remainder of this section. However, retrieving incorrect inference results within the deadline can be extremely harmful for a special purpose functionality. For this reason, we need to define a measure of inference confidence to be incorporated into the overall definition of goal effectiveness.

3.2.3. On the Use of Entropy to Define the Goal Value

We now introduce the metric reflecting inference accuracy, which will be then used to define the goal value. As we are dealing with a classification task, the actual performance metric could be the one of correct pattern classification rate. However, to evaluate this metric, the ground truth, i.e., the true label, would need to be always available, which is not the case, in general. In a practical setting, a metric suitable for evaluating classification tasks as they emerge needs to be characterized by the following features: (i) it must be measurable online without any ground truth; (ii) it should reflect, in the best possible way, the value of the goal, i.e., the classification accuracy. A possible metric fulfilling these requirements is the entropy computed on a posteriori probabilities, which are the typical output of any discriminative or generative classifier, such as a DNN [25,55,56]. In particular, the output of a discriminative classifier can be written as a vector $\mathbf{p} = [p_1, \dots, p_L]^H$ of probabilities, with each probability being associated with one out of the L possible labels of the data set under investigation. Obviously $\sum_{l=1}^L p_l = 1$ holds true. Given \mathbf{p} , the entropy associated to a classified pattern b is

$$H_b = -\mathbf{p}^H \log(\mathbf{p}). \quad (12)$$

The entropy is a scalar measure, which can be interpreted as the classification confidence of an ML model in classifying a pattern. The lower the entropy, the more confident the classifier is. Namely, for a data set with L labels, the worst case for a single pattern classification is $p_l = \frac{1}{L}, \forall l = 1 \dots, L$ (i.e., throwing an L facets dice), which translates into the maximum entropy $H_{\text{max}} = \log(L)$; also, noting that $\lim_{p \rightarrow 0^+} p \log(p) = 0$, and that $p \log(p) = 0$ if $p = 1$, the minimum value attained by the entropy in case of a completely sure classifier is 0. Therefore, the entropy is always a bounded metric $0 \leq H \leq H_{\text{max}} = \log(L)$. Since we are considering inference on batches of data (cf. (10)), we can define the *average entropy* H_t on the batch classified during slot t as

$$H_t = \frac{1}{N_t} \sum_{b \in \mathcal{B}_t} H_{b,t}, \quad (13)$$

where \mathcal{B}_t denotes the set of patterns in the batch and $N_t = \text{card}(\mathcal{B}_t)$ (cf. (10)). For simplicity, and for the sake of coherence with the goal-value definition in Section 2, we will use, as goal value, the negative relative average entropy increase (referred to as *NREI* in the sequel), with respect to the its minimum value, attained on a validation subset of the original data set, i.e., without any communication impairment, i.e., interference from DO user (It can be, for instance, assumed that model training is performed using a radio bandwidth (of same size as in inference phase) exclusively used by the GO system.) This choice is dictated by the fact that, even for the original data set, the average entropy is not actually 0, but it attains a minimum value that depends on the training phase. Therefore, denoting by H_{min}

the minimum average entropy over the considered data set, we write the NREI (i.e., the goal value) as follows:

$$\Theta_t = -\frac{H_t - H_{\min}}{H_{\min}}. \quad (14)$$

Then, the higher the absolute value of the NREI, the lower the goal value is, and vice versa. A natural question arises on the dependence of this goal value on wireless performance, e.g., on the PER. In this paper, since we consider the PER as a source of classification entropy increase/decrease, we make the following principled and mild assumptions on entropy and accuracy, which we validate by experimental results.

Assumption 1. *The classification entropy is a monotonic non-decreasing function of the PER.*

This assumption, despite not being supported by a mathematical proof, is based on the rationale that a more severely degraded version of the inferencing input data (i.e., distorted input patterns received by AP_g) generates higher “confusion” in the classifier, thus increasing the classification entropy.

Assumption 2. *The classification accuracy is a monotonic non-increasing function of the average entropy.*

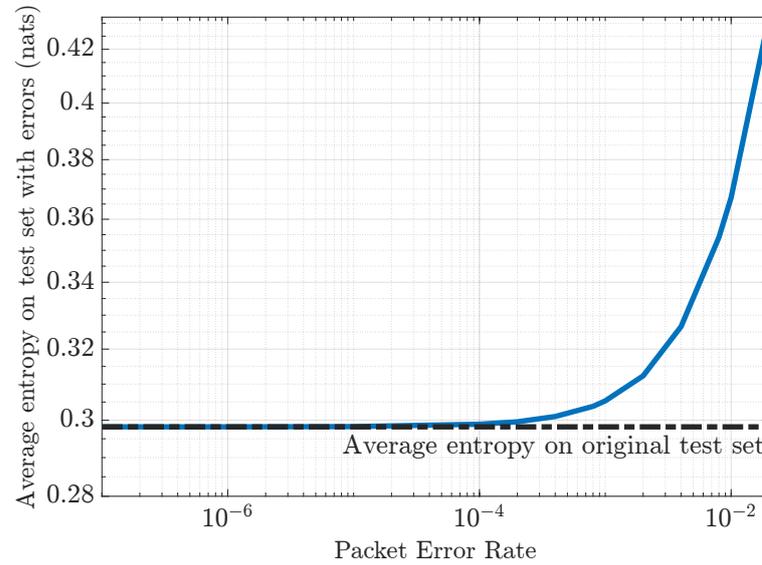
Assumption 2 is also not supported by a mathematical proof and can fail in some specific cases (e.g., a very different data distribution in a test set). However, it is mild and valid in many operating conditions. In addition, the cross-entropy is the typical loss for DNN training. In Figure 2, we validate the two assumptions on the CIFAR-10 data set [57], whose detailed description, along with the one of the trained CNN architecture, is provided in Section 3.4.2 to enlighten the reader. In particular, Figure 2a shows the average entropy over the test set, as a function of the PER, while Figure 2b shows the test accuracy as a function of the average entropy. Finally, for the sake of complete definition of goal effectiveness, we remind that the end-to-end inferencing delay is given by (11).

Now, we can formalize the goal and its effectiveness for the investigated system, i.e., the edge inference service, coexisting with a legacy communication service. The goal is to obtain inference results within a maximum delay D_{\max} from the time instant the classification request is issued, with goal value (cf. (14)) higher than a predefined threshold Θ_{th} (in this paper, with reference to a batch of patterns, we assume that the time instant of batch-based classification task generation coincides with the start of batch transmission in the uplink; this assumption can be easily generalized and will be investigated in future works). Formally, the goal is achieved if the event $\{D_{\text{tot},t} \leq D_{\max}\} \cap \{\Theta_t \geq \Theta_{\text{th}}\}$ occurs, or, equivalently, a *goal outage* is represented by one of the following events: (i) $\{D_{\text{tot},t} > D_{\max}\} \cup \{\Theta_t \geq \Theta_{\text{th}}\}$, (ii) $\{D_{\text{tot},t} \leq D_{\max}\} \cup \{\Theta_t < \Theta_{\text{th}}\}$ or (iii) $\{D_{\text{tot},t} > D_{\max}\} \cup \{\Theta_t < \Theta_{\text{th}}\}$. Then, we can define the goal effectiveness as the probability of achieving the goal:

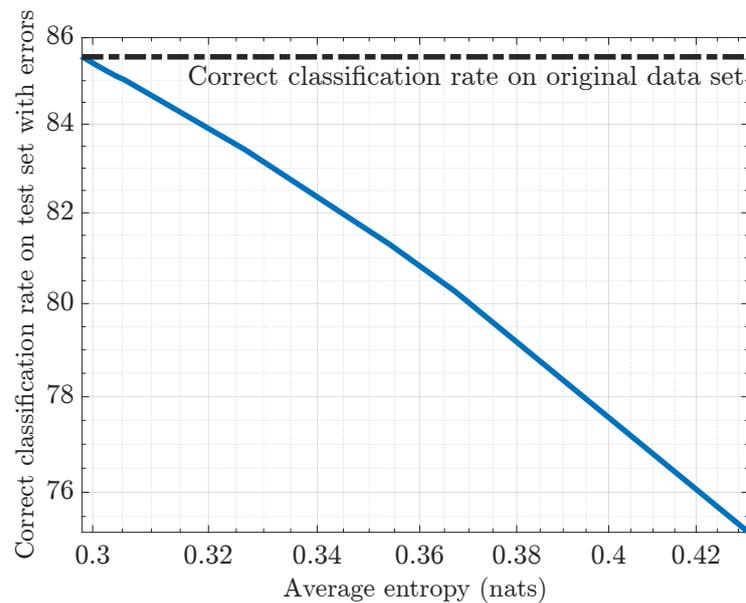
$$\mathcal{E}_g = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\{ \mathbf{1}_{\{\Theta_t \geq \Theta_{\text{th}}\}} \cdot \mathbf{1}_{\{D_{\text{tot},t} \leq D_{\max}\}} \right\}. \quad (15)$$

Now, let us note that the data rate in (8) is a monotonic increasing function of the target PER γ_t , i.e., a higher PER tolerance allows UE_g to transmit data at higher speed, which also means lower E2E delay (cf. (10) and (11)). This straightforward yet fundamental behavior and the goal-effectiveness definition in (15) come with non-trivial consequences, among which we identify the following: the edge inference system can be either limited by the inference entropy or by the E2E delay, depending on the selected target PER and received interference by the DO system (i.e., wireless performance), as well as on edge processing delay (i.e., computation performance). Indeed, due to Assumption 1, lower PER leads to better inference performance (from a classification confidence point of view), while it could be highly detrimental from an E2E delay, and, therefore, classification timeliness perspective

(cf. (8)). As a consequence, it is not necessarily more convenient to transmit classification input data with ultra high communication reliability (e.g., target PER below 10^{-7}), but rather with the minimum level of communication reliability that *jointly* guarantees the *target level of classification confidence and its timeliness*. We argue that, based on the definition in (15), a lower goal effectiveness could be achieved as a result of ultra-reliable yet slower wireless communication, which is the purpose of the next section.



(a) Average entropy vs. PER



(b) Accuracy vs. entropy

Figure 2. Validation of Assumptions 1 and 2.

The aim of the goal-oriented optimization approach is to find the right balance between the two measures, to attain the desired goal effectiveness, entailing both inference timeliness and inference confidence (goal value). So far, we have not formalize the goal cost of the scenario under our investigation.

3.3. The DO User Data Rate Loss as Goal Cost

Let us recall that our aim is to assess performance in a scenario in which a GO system coexists with a DO system. In the considered scenario, the DO user, UE_d , uploads data

(e.g., a video stream) with the aim of maximizing its data rate. However, the system is cooperative, and the DO user aims to maximize its data rate without preventing the GO user from attaining the desired goal effectiveness. Denoting by $\mathbf{w}_d \in \mathbb{C}^{M_d \times 1}$ the combining vector of AP_d , the received SINR at AP_d at time t reads as follows:

$$\text{SINR}_{d,t} = \frac{|\mathbf{w}_{d,t}^H \mathbf{h}_{dd,t}|^2 P_{d,t}}{N_0 W + |\mathbf{w}_{d,t}^H \mathbf{h}_{dg,t}|^2 P_{g,t}}, \quad (16)$$

and the data rate can be approximated by the Shannon formula, as we do not need to assess the PER performance for the DO system (however, generalizing this would be straightforward):

$$R_{d,t} = W \log_2(1 + \text{SINR}_{d,t}). \quad (17)$$

Denoting by $\mathcal{R}_{d,\max}$ the maximum achievable average data rate by the DO system in the absence of the GO one and with transmission power $P_d = P_{d,\max}$, where $P_{d,\max}$ stands for the maximum transmission power of UE_d , the goal cost can be defined as the *average relative UE_d data rate loss*, i.e., it can be written as follows (cf. (1)):

$$c = \frac{\mathcal{R}_{d,\max} - \mathcal{R}_d}{\mathcal{R}_{d,\max}}. \quad (18)$$

The lower the DO user data rate is, the higher the goal cost is. In other words, the price to make the GO user achieve its goal with target goal effectiveness is paid by the DO user through its achievable uplink data rate. In the sequel, we assume that the DO user always has backlogged traffic, i.e., it continuously transmits and interferes with the GO system. This can be easily generalized.

Now, we have all elements describing goal effectiveness and goal cost in the investigated scenario. In the next section, to gain insights on relevant parameters to be controlled and optimized in the proposed system, we will first present a performance assessment, obtained through Monte Carlo simulations. Then, based on the acquired messages, a goal-oriented problem formulation, along with its solution, will be presented in Section 4. We believe that this gentle introduction will help the reader understand all ingredients of the system, and also the reasoning behind the goal-oriented approach and the corresponding proposed optimization problem.

3.4. Evaluation of the System without Optimization

In this section, we evaluate the performance of the overall system, in which GO and DO systems coexist. To do so, Monte Carlo experiments are conducted, considering different realizations of wireless channels, computation delays and pattern inference requests (i.e., number of patterns generated at each time slot). Results are obtained from $T = 50,000$ independent realizations of such parameters. The following assumptions hold for communication, computation, and inference settings.

3.4.1. Wireless Communications Assumptions

We consider two users (i.e., UE_g and UE_d), two APs (i.e., AP_g and AP_d) and an MEH, as in Figure 1. UE_g , AP_g , UE_d and AP_d are placed at $[5, 0]$, $[5, 20]$, $[8, 0]$ and $[8, 20]$, respectively. Both UEs are equipped with a single antenna, and both APs employ a uniform linear array of $M_g = M_d = 8$ elements. Maximal ratio combining is employed for the receive filters at both AP_g and AP_d . Channel coefficients are generated as in (6), with path loss β_{ij} with exponent 4, while Rice fading with factor $K = 3$ is considered. A different channel realization is extracted for each experiment from the described distribution. To evaluate the performance, we assume the transmit power of UE_g to be fixed to $P_g = 100$ mW, while we vary the transmit power of UE_d , choosing it in $[0, P_{c,\max}]$ W, with 500 evenly spaced points, and $P_{c,\max} = 200$ mW, in order to explore the performance as a function of UE_d data rate (in each simulation, we will specify the selected value whenever needed). The carrier frequency is $f_c = 28$ GHz (<https://www.bmwk.de/Redaktion/EN/Publikationen/Digitale-Welt/>

[guidelines-for-5g-campus-networks-orientation-for-small-and-medium-sized-businesses.pdf](#) accessed on 20 July 2023), with total bandwidth $W = 1$ GHz, fully shared among UE_g and UE_d , while the noise power spectral density is fixed to $N_0 = -174$ dBm/Hz, with a 3 dB noise figure at the receiver. The transmission of UE_g is organized in packets, each of equal size of 32 Bytes. The PER is chosen from the vector $\gamma = [10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 4 \times 10^{-4}, 8 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 4 \times 10^{-3}, 8 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}]$ to explore performance in terms of communication reliability. Therefore, at time t , given a PER γ_t , the transmitted packets are considered to be incorrectly received with probability γ_t chosen from the vector (and specified for each simulation whenever needed). When a packet is not correctly received, no retransmission is requested, and the bits within the packet payload are randomly chosen.

3.4.2. Data Set and Inference Assumptions

We evaluate the performance on the CIFAR-10 data set [57], which consists of 32×32 pixel RGB images, with 50,000 patterns in the training set and 10,000 patterns in the test set. We resize each image to have 64×64 pixel RGB images, representing each of the 3 base colors with 32 bits. We assume that a pre-trained CNN is pre-onboarded at the MEH and ready to provide classification results. We consider batches \mathcal{B}_t of equal size of $N_t = 20$ patterns (cf. (10) and (13)) for each realization. Therefore, considering the packet size of 32 Bytes and the image size, the total number of packets transmitted for each experiment (i.e., time slot) is $\frac{64^2 \times 3 \times 32 \times 20}{256} = 30,720$. We assume that a state-of-the-art model (<https://appliedmachinelearning.wordpress.com/2018/03/24/achieving-90-accuracy-in-object-recognition-task-on-cifar-10-dataset-with-keras-convolutional-neural-networks/> accessed on 20 July 2023) is pre-trained and pre-uploaded in the MEH. The model is trained using Matlab[®]. For simplicity, we also assume a fixed blocklength $n_g = 64$ Bytes.

3.4.3. Computation Delay Assumption

In the absence of a model for the computing delay, we have empirically built a delay distribution from real-world experiments. The inference runs on a GPU NVIDIA[®] Tesla[®] V100. The CPU characteristics are the following: Intel[®] Xeon[®] Gold 6244 CPU@ 3.6 GHz, with 4 cores and 8 GB of memory. Then, we run batch inference for $T = 50,000$ independent realizations, we save the computing time and we retrieve a computing delay distribution, from which we can extract a random realization when evaluating performance. For the interest of the reader, the estimated probability density function of the computing delay is shown in Figure 3, for all values of PER, as (a priori) the distribution may change as a function of the PER. However, from the figure, it can be noted that the distribution is stable over PER values, showing only slight variations, probably caused by fluctuations in the CPU usage and background processes. Therefore, in the sequel, we will consider the computation delay distribution as independent from the target PER.

We will now present the results in terms of goal effectiveness, as per its definition in (15). Let us recall that the goal effectiveness generally entails the goal value and other system constraints. For instance, in the case of edge inference, the goal value is represented by the NREI (cf. (14)), and the other system constraint is represented by the E2E delay (cf. (11) and (10), Figure 3). However, for the sake of smooth exposition, we will first present two separate results, i.e., the goal effectiveness by only taking into account the goal value—NREI (c.f. (14)), and the goal effectiveness by only taking into account the delay, respectively. We believe that this gentle introduction will help us with commenting on the results, and the reader with comprehending the conveyed message, via a clear explanation of the different sources of goal outages. The two notions of goal effectiveness will be then merged in Section 3.4.6 as a final result to coherently consider the complete definition in (15). Afterwards, an optimization problem will be proposed in Section 4.

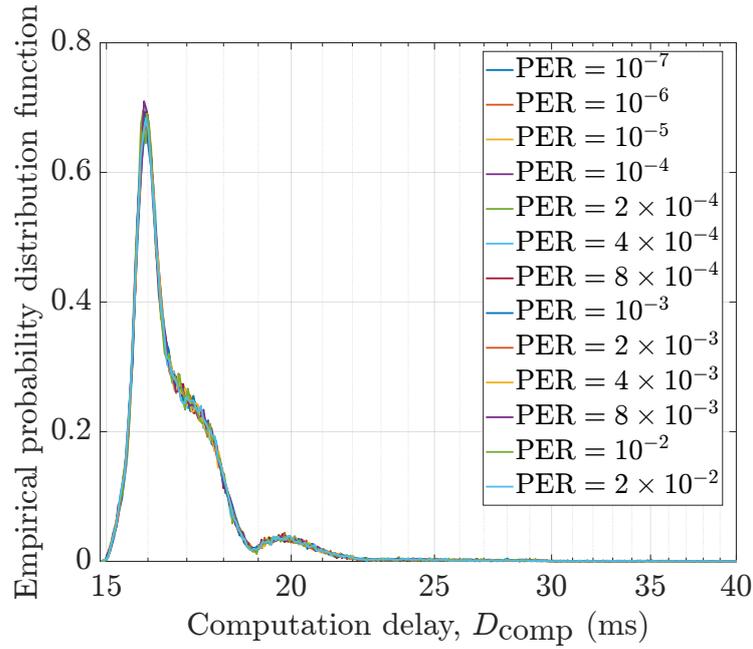


Figure 3. Empirical probability density function of computing delay, estimated through real experiments.

3.4.4. The Goal Effectiveness from the Goal-Value Perspective

Let us evaluate the performance of goal effectiveness, only from the point of view of the goal value, i.e., the first term on the right-hand side of (15) (the inference reliability—negative relative entropy increase—NREI). In other words, we consider a received batch result with goal value higher than the predefined threshold, as a goal achievement, even if results are issued after the deadline D_{\max} . Let us first notice that the goal value, as per its definition in (14), is only affected by the PER, i.e., the errors generated by the wireless communication between the GO user and the GO AP. Then, in Figure 4, we show the goal effectiveness related to the goal value as a function of the PER for different thresholds Θ_{th} . First, we can notice how the probability of the goal value being above a threshold decreases as the PER increases and as more errors occur throughout the uplink communication phase. This result is also in line with Assumption 1 (Figure 2a). In addition, by increasing the threshold Θ_{th} (recall that the average NREI assumes non-positive values), performance degrades, as expected.

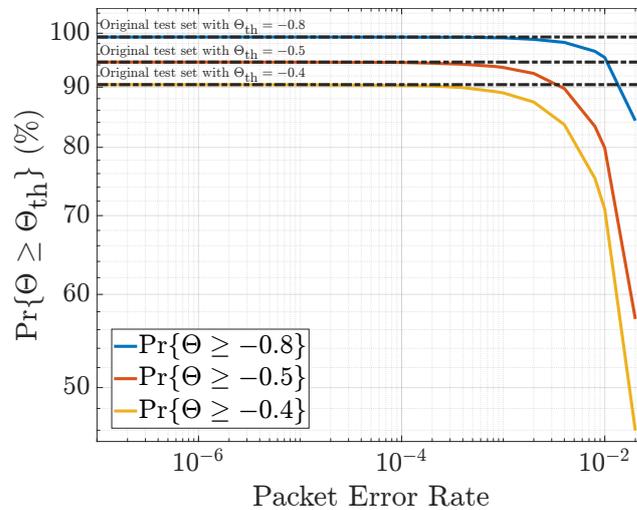


Figure 4. Probability of goal value (NREI cf. (14)) being above a predefined threshold, as a function of PER, for different threshold values.

As already mentioned, this result does not take into account the delay, which is a fundamental part of the goal (i.e., the edge inference service). In the following, we consider the goal effectiveness from the delay perspective then merge the two perspectives in Section 3.4.6.

3.4.5. The Goal Effectiveness from the Delay Perspective

First, it is worth noting that the E2E delay depends on the computation delay, the PER and the interference (cf. (8) and (10)). In Section 3.4.3 (Figure 3), we empirically illustrated the independence between PER and computation delay. On the other hand, the communication delay is highly dependent on PER and received interference. Indeed, at fixed channel conditions and GO user transmit power P_g , by increasing the DO user's transmit power P_d , the data rate of the latter increases, thus generating more interference to the GO system, which causes higher delay in the case of fixed-target PER. To show the effect of the E2E delay on the effectiveness, we consider a delay threshold $D_{\max} = 50$ ms, and we plot, in Figure 5, the probability for a batch to be classified by the deadline as a function of the DO user data rate loss (i.e., the goal cost in (18)) for a subset of target PER $[10^{-7}, 10^{-4}, 10^{-3}, 10^{-2}]$ (for ease of readability). This is obtained by increasing the DO user transmit power P_d by $[0, 200]$ mW. As we can notice, contrarily, as before, lower PER leads to lower effectiveness (as the effectiveness only entails delay, i.e., the GO user does not care about receiving high entropy results, provided that they are received within the predefined E2E delay). This result is due to the strong dependence of the communication delay (and thus the E2E delay) on PER and interference. Indeed, as an example, let us show, in Figure 6, the empirical probability density function of the GO system E2E delay (communication and computation) for the left-hand point of Figure 5, i.e., for a single value of $P_c = 200$ mW. As we can notice, the distribution of the delay experiences longer tails for lower PER, as expected from (8). From the last results, we can easily conclude that:

- From a goal-value perspective, the goal effectiveness is only affected by the PER (although not strongly, depending on the PER value), and a higher PER (i.e., lower communication reliability) leads to lower (partial) goal effectiveness; also, from $\gamma = 10^{-7}$ to $\gamma = 10^{-3}$, stable performance is experienced.
- From an E2E-delay perspective, the goal effectiveness is affected by the PER and the DO user interference, and higher PER (i.e., lower communication reliability) leads to higher (partial) goal effectiveness.

The goal of this work is to analyze and optimize performance by taking into account the overall definition of a goal (and goal effectiveness), entailing the goal value and the end-to-end delay but also the goal cost. Therefore, in what follows, we finally present the results in terms of goal effectiveness, as per its overall definition in (15), together with the corresponding goal cost.

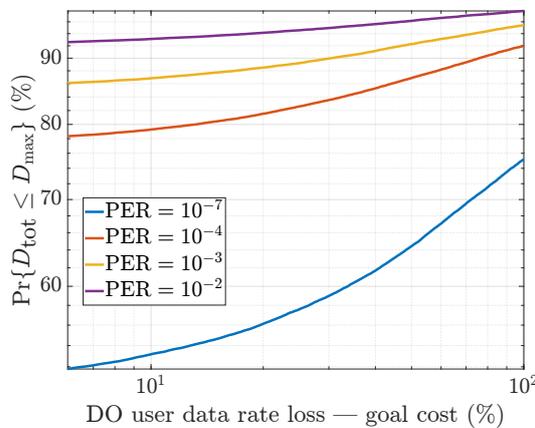


Figure 5. Probability of guaranteeing the delay constraint vs. goal cost (DO user relative data rate loss).

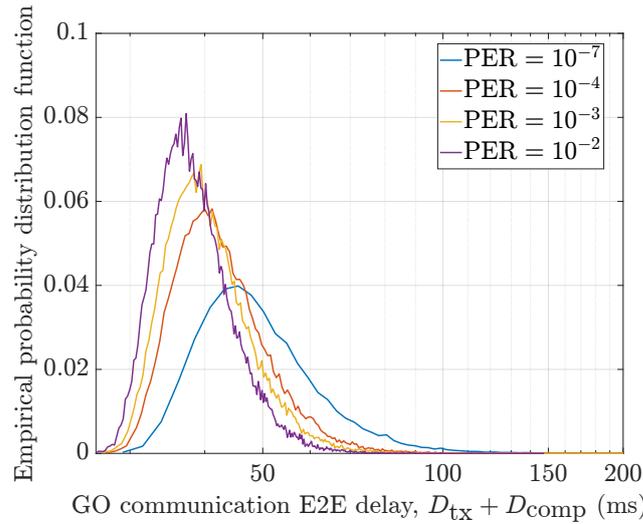


Figure 6. Empirical probability density function of E2E delay for goal cost $\mathcal{C} \approx 6\%$ (left-hand points of Figure 5).

3.4.6. The Goal Effectiveness and its Dependence on PER, Interference and Goal Cost

The goal effectiveness in (15) depends on two performance indicators: (i) the goal value (represented by the NREI), which, in our setting, directly and exclusively depends on the PER, and (ii) the E2E delay, which depends on the PER, the interference caused by the DO user and the remote computation delay. As a first joint result, in Figure 7, we show heat maps representing the goal effectiveness (i.e., the goal effectiveness is represented by the different colors, from dark blue to yellow) as a function of PER (y -axis) and maximum delay threshold D_{\max} (x -axis) for a fixed DO user data rate (i.e., fixed $P_d = 200$ mW) and for different goal value thresholds in three different plots, i.e., $\Theta_{\text{th},1} = -0.8$, $\Theta_{\text{th},1} = -0.5$ and $\Theta_{\text{th},1} = -0.4$, shown in Figure 7a–c, respectively (we remind that a higher threshold indicates a stricter constraint in the sense of goal achievability). Note that the heat maps are interpolated a posteriori for better visualization, while results have been obtained for the sub-selected PER and D_{\max} values. In these figures, we also show different goal-effectiveness thresholds (i.e., $\mathcal{E}_{g,\text{th}} = 0.7$, $\mathcal{E}_{g,\text{th}} = 0.8$, $\mathcal{E}_{g,\text{th}} = 0.9$), through contours plots. All the points interior to these contour plots represent a *goal feasibility region* for each respective goal-effectiveness constraint, i.e., all the combinations of (fixed) PER and end-to-end delay, whose corresponding goal effectiveness exceeds a predefined threshold. However, these regions are subject to the fact that there is no adaptation of transmit power and target PER across time; whereas, as we will show in the sequel, much better results can be obtained in the optimized setting. From these plots, we can make the following considerations:

- The goal effectiveness increases as the delay threshold (y -axis) increases (for each fixed target PER), while it does not necessarily decrease as a function of the PER, as expected and shown before in the disjoint plots, creating *goal effectiveness feasibility regions*, whose surface depends on $\mathcal{E}_{g,\text{th}}$.
- The goal effectiveness feasibility regions shrink as the goal value (NREI) threshold increases (see the difference between surface extensions in Figure 7a–c—looking at Figure 7c, we can even notice that no combination of fixed PER and D_{\max} guarantees goal effectiveness above 0.9 for $\Theta_{\text{th}} = -0.4$ in this setting).
- Given a goal-effectiveness requirement, there always exists a minimum E2E delay threshold guaranteeing feasibility; whereas, below this threshold, it is infeasible (for any PER) to guarantee the requirement (examples of this point are shown by the red arrows in the figures). Moreover, each target PER experiences a different minimum D_{\max} that can be guaranteed. The lower the PER is, the higher the minimum feasible delay is.

- As the delay threshold decreases, the feasible region in terms of PER also shrinks, i.e., with a lower delay constraint, higher PER values are needed to guarantee effectiveness; however, this is not always feasible due to the goal value constraint (see, e.g., Figure 7c).
- As a consequence of the previous remark, as the PER decreases (y -axis), the minimum delay threshold to guarantee a target goal effectiveness increases, i.e., to guarantee lower PER (reliable communication), UE_g needs more time for offloading, resulting in more frequent delay outages.

Let us now focus on Figure 7d–f, in which we plot the goal effectiveness as a function of DO user data rate loss in (18), i.e., the goal cost (y -axis) and the PER (x -axis), for a fixed $D_{\max} = 45$ ms, and different goal value (NREI) thresholds across figures, namely the same as the previous results. In addition, a contour plot representing the 80% goal effectiveness feasibility region is shown in all figures. These plots are obtained by also varying the DO user transmit power (within the range $[0, 200]$ mW). From these results, we can appreciate the goal feasibility region, also as a function of the goal cost (i.e., the aim of this paper), and we can draw the following conclusions:

- Again, the goal effectiveness feasibility region is a surface, i.e., there are multiple solutions guaranteeing the goal-effectiveness constraint.
- While the above consideration holds, there exists a minimum goal cost solution, i.e., the minimum cost needed to achieve the target goal effectiveness. The latter is the lowest point of the contour plots representing the effectiveness thresholds and is represented by the black horizontal dashed lines in each plot.
- By increasing the goal value threshold (i.e., across different figures), the feasibility region shrinks as before and, as an additional observable effect, the minimum goal cost increases (e.g., above 60% of DO user data rate loss in Figure 7e). In other words, the stricter the constraint in terms of goal value is, the higher the minimum goal cost able to guarantee effectiveness is.
- In certain conditions (see Figure 7f), desired values of goal effectiveness are not attainable (e.g., goal effectiveness above 80%)

Interestingly, these results relate the performance of a legacy (DO) and a GO communication system interfering with each other, showing non-trivial outcomes, which suggest that higher communication reliability does not necessarily imply higher goal effectiveness. Therefore, they represent the basis to formulate a goal-oriented optimization problem in the next section. Indeed, the aim of a goal-oriented resource orchestration framework is to move within the goal effectiveness feasibility region, possibly finding the lowest cost in such region.

Then, from these results and their corresponding conclusions, it is straightforward to formulate a goal-oriented resource allocation problem involving the variables that mostly affect goal effectiveness and goal cost: (i) the PER of GO communication and (ii) the DO user transmit power, in this work. Other variables can be taken into account, which represents further research directions on this topic.

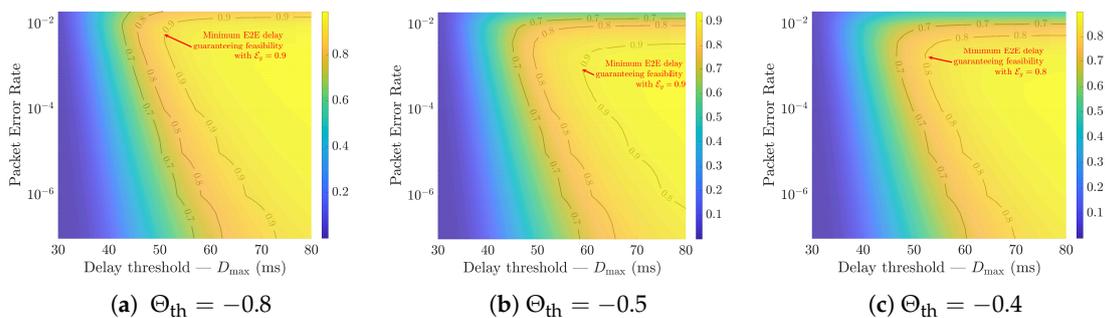


Figure 7. Cont.

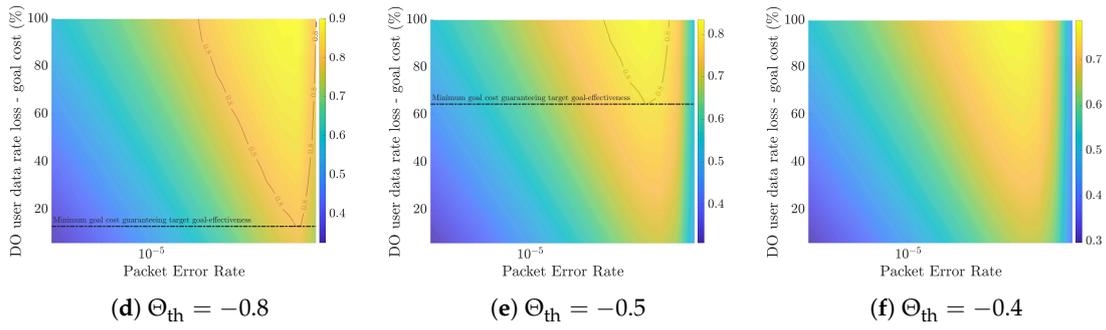


Figure 7. (a–c) Goal-effectiveness heat maps as a function of GO communication PER and E2E delay constraint for fixed goal cost and different goal value thresholds Θ_{th} ; (d–f) goal-effectiveness heat maps as a function of goal cost (relative DO user data rate loss) and target PER of the GO system for fixed GO communication E2E delay threshold (45 ms) and different goal value thresholds Θ_{th} .

4. Problem Formulation and Solution

The aim of this section is to propose a resource allocation policy able to minimize the UE_d data rate loss (cf. (18)), subject to a goal-effectiveness constraint of the GO system. Of course, minimizing (18) is equivalent to maximizing the average data rate of UE_d . Then, following the general formulation in (5), the edge inference problem can be formulated as follows:

$$\max_{\{\Psi_t\}_t} \mathcal{R}_d := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_{d,t}\} \quad (19)$$

subject to

$$(a) \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \mathbf{1}_{\{D_{tot,t} \leq D_{max}\}} \right\} \geq \mathcal{E}_{g,th},$$

$$(b) P_{d,t} \in \mathcal{P}_d, \forall t, \quad (c) \gamma_t \in \Gamma_g, \forall t,$$

where $\Psi_t = [P_{d,t}, \gamma_t]$ is the action set, i.e., the optimization variables, involving DO user transmit power and target GO user PER. The constraints of the problem have the following meaning: (a) the goal effectiveness of the GO system is higher than a predefined threshold; (b) the instantaneous transmit power of UE_d is chosen from a discrete set \mathcal{P}_d involving a minimum value of transmission power (0 in this case) and a maximum value being equal to $P_{d,max}$; (c) the target PER of the GO system is chosen from a discrete set Γ_g . In addition, we make the following assumptions:

1. The goal is achievable, i.e., problem (19) is feasible.
2. The optimization is performed at the MEH, which is provided with the needed connect-compute instantaneous information, as specified here below.
3. All effective channels (i.e., including the receive filters) are perfectly known instantaneously, while their statistics are unknown in advance.
4. The computation delay at the current time slot is estimated and known with high confidence, i.e., we assume the computation delay is known at time t .
5. The GO user has no buffered data, but it is able to accept all data patterns generated at time t , to be transmitted to the MEH.
6. The DO user always has backlogged traffic, i.e., it continuously transmits and interferes with the GO system
7. All thresholds (delay, entropy, effectiveness) are known in advance, i.e., they are requested as part of a service-level agreement.

Problem (19) is difficult to solve due to the lack of distribution knowledge of wireless channels (i.e., only instantaneous realizations are observed) and computation delays at the MEH. Therefore, we hinge on the tools of Lyapunov stochastic network optimization to solve the problem on a per-slot basis, with instantaneous observations of context parameters. To this end, let us first introduce the concept of *virtual queue* [58], whose aim is to keep track of constraint violations and take specific actions to drive the system towards desired operating modes. In particular, given a long-term constraint written as

$$\mathcal{X} := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{x_t\} \leq x_{\text{th}}, \quad (20)$$

we can define an associated virtual queue that evolves as follows over successive time slots:

$$Z_{t+1} = \max(0, Z_t + x_t - x_{\text{th}}). \quad (21)$$

The evolution of this mathematical model is straightforward: the size of the virtual queue grows whenever the constraint is violated, and it decreases otherwise. Interestingly, as easily proved in ([58], Section 4.4) constraint (20) is guaranteed if the associated virtual queue is *mean-rate stable*, i.e.,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}\{Z_T\} = 0. \quad (22)$$

To ensure the mean-rate stability of the virtual queue, it is sufficient to guarantee that the so-called *Conditional Lyapunov Drift (CLD)* is bounded by a finite constant at each slot. Therefore, let us introduce the CLD by first defining the Lyapunov function [58]

$$L(Z_t) = \frac{1}{2} Z_t^2, \quad (23)$$

which is a measure of the congestion state of the system in terms of the defined virtual queue(s). From (23), the CLD is defined as follows:

$$\Delta_t(Z_t) = \mathbb{E}\{L(Z_{t+1}) - L(Z_t) | Z_t\}, \quad (24)$$

i.e., it is the expected variation of the Lyapunov function over two successive time slots. Guaranteeing that $\Delta_t(Z_t) \leq B$, with B a finite constant, also ensures the mean-rate stability of Z_t , and therefore constraint (20). Now, for the virtual queue defined in (21), it is easy to prove that

$$\frac{Z_{t+1}^2 - Z_t^2}{2} \leq \frac{(x_{\max} - x_{\text{th}})^2}{2} + Z_t(x_t - x_{\text{th}}), \quad (25)$$

where x_{\max} is the (finite by hypothesis) maximum value that x_t can take, given system design constraints. Then, one can write the following upper bound ([58], Equation (4.47)):

$$\Delta_t(Z_t) \leq \frac{(x_{\max} - x_{\text{th}})^2}{2} + \mathbb{E}\{Z_t(x_t - x_{\text{th}}) | Z_t\}. \quad (26)$$

Following stochastic optimization arguments [58], it is sufficient to remove the expectation and minimize the CLD upper bound in (26) in a per-slot basis to guarantee the mean-rate stability of Z_t and, as a consequence, constraint (20), under the assumption of i.i.d. realizations of context parameters (an assumption that can be relaxed in some cases). However, in this way, no importance is assigned to the objective function of the problem, i.e., the goal cost in this case. To take the latter into account, denoting by C_t the goal cost at time t , one can write the so-called *drift-plus-penalty (DPP)* function [58]:

$$\Delta_p(Z_t) = \mathbb{E}\{L(Z_{t+1}) - L(Z_t) + \Omega \cdot C_t | Z_t\}, \quad (27)$$

where Ω denotes a scalar hyper-parameter (the only one) controlling the trade-off between (original) constraint guarantees and objective function value minimization. The DPP is an augmented version of the CLD, and it penalizes high values of the objective function, e.g., excessive usage of resources in a network. Interestingly, by bounding the DPP, one

can derive a similar result on virtual queue stability and long-term constraint guarantees, i.e., if $\Delta_p(Z_t) \leq B, \forall t$, the same result applies, with the following difference: by increasing the trade-off hyper-parameter Ω , the value of the objective function decreases at the cost of longer convergence time of the virtual queue and its average value ([58], Theorem 4.8). Finally, to adapt the above analysis to problem (19), we can define the following virtual queue evolution for constraint (a) of problem (19):

$$Z_{t+1} = \max\left(0, Z_t - \mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \mathbf{1}_{\{D_{tot,t} \leq D_{max}\}} + \mathcal{E}_{g,th}\right). \quad (28)$$

For virtual queue Z_t in (28), we can write (cf. (25)):

$$\begin{aligned} \frac{Z_{t+1}^2 - Z_t^2}{2} &\leq \frac{(1 - \mathcal{E}_{g,th})^2}{2} \\ &\quad - Z_t \left(\mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \mathbf{1}_{\{D_{tot,t} \leq D_{max}\}} - \mathcal{E}_{g,th} \right), \end{aligned}$$

which leads to the following DPP upper bound (recalling the objective function of (19)):

$$\begin{aligned} \Delta_p(Z_t) &\leq B_1 + \mathbb{E}\left\{ -Z_t \left(\mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \mathbf{1}_{\{D_{tot,t} \leq D_{max}\}} - \mathcal{E}_{g,th} \right) \right. \\ &\quad \left. - \Omega \cdot R_{d,t} | Z_t \right\}, \end{aligned} \quad (29)$$

with $B_1 = \frac{(1 - \mathcal{E}_{g,th})^2}{2}$. Finally, by minimizing (29) in a per-slot basis, we obtain the following instantaneous problem (we omit the constant terms):

$$\min_{\Psi_t} -Z_t \mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}} \cdot \mathbf{1}_{\{D_{tot,t} \leq D_{max}\}} - \Omega \cdot R_{d,t} \quad (30)$$

subject to

$$(a) P_{d,t} \in \mathcal{P}_c, \quad (b) \gamma_t \in \Gamma_g.$$

Hinging on the theoretical findings in ([58], Th. 4.5 and Th. 4.8), under the i.i.d. assumption of context parameters (i.e., wireless channels and remote computing delay), by solving (30) in each time slot, the mean-rate stability of the virtual queue is guaranteed (i.e., constraint (a) of (19) is met). Moreover, asymptotic optimality is achieved as the trade-off parameter Ω increases, at the cost of higher average virtual queue value and convergence time. Finally, thanks to the concept of Γ -additive approximation, non-exact solutions are allowed in expected sense, provided that the solution is within a bounded value Γ from the infimum of all possible solutions, with an impact on the optimality performance of the algorithm that depends on Γ . Then, the next step is to solve problem (30). Problem (30) is an integer program, however, on a limited feasible set with cardinality $|\mathcal{P}_c| \times |\Gamma_g|$. As such, we will simply solve it through an exhaustive search over the feasible set. This is a typical outcome of Lyapunov stochastic optimization, thanks to the decoupling of the long-term problem into a sequence of (simpler) problems, based only on instantaneous observations of context parameters and properly defined state variables (virtual queues). The virtual queue (which evolves over time) and the parameter Ω drive the trade-off between UE_d data rate loss (i.e., the goal cost), UE_g goal effectiveness and convergence time.

4.1. Solution of the Instantaneous Problem

Let us notice that the first indicator function in (30) ($\mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}}$) is strongly dependent on the choice of the PER (equivalently, in practice, on the specific MCS employed for GO communication at time slot t). It should be noted that, in general, there is no known function relating PER and entropy relative increase (goal value). In addition, differently from the one related to the delay (which is known if system state is perfectly known at the MEH—as we assume here), the actual value of such indicator variable is made available by the server only *after the inference takes place*, i.e., after the computation phase, which is obviously too late to make a decision. Therefore, the decision has to be made on a “best

guess" of what the outcome could be. For this reason, as an approximation to solve (30), we replace the first indicator function with its expectation, i.e., the probability of the goal value exceeding the threshold. Nevertheless, this approximation is not enough to solve the problem, as a model linking the PER and this probability is generally not available in advance. One possible solution is to exploit model-free optimization tools, such as DRL [59]. However, as the scope of this paper (and, in particular, of this section) is to provide a contribution to the coexistence of GO and DO systems, we will rely on a look-up table built on a validation set. Specifically, using a validation set from the considered input pattern set, we build a look-up table linking PER at AP_g and the probability of the goal value being above the predefined threshold, to be used during operation time to select the desired PER, based on current virtual queue states and context parameter realizations.

Once the first indicator function in (30) (i.e., $\mathbf{1}_{\{\Theta_t \geq \Theta_{th}\}}$) is replaced with its expected value (i.e., the probability of the event) and the look-up table is built, the remaining part of the problem can be *optimally solved* via an exhaustive search over the involved variables. The overall procedure to dynamically select DO user transmit power and GO system PER is described in Algorithm 1.

Algorithm 1: Goal-oriented resource allocation

At each time slot t :

1. Observe wireless channel realizations, computation resources and virtual queue states;
 2. Solve (30) as described in Section 4.1
 3. Observe the real goal value outcome at the output of the classifier (14), and update the virtual queue Z_t as in (28)
 4. Go to next time slot $t + 1$
-

The main challenge of the proposed method is the need of modeling the whole system, including the inference performance as a function of the PER. In more practical scenarios, such a system modeling requirement can be bypassed through the exploitation of data-driven techniques, such as RL. By applying such an approach, the system can adapt through exploration and exploitation of previously performed actions to find the right balance towards an optimized policy. The latter investigation is left for future works, while, in this paper, we make use of a model to give insights on the achievable system performance.

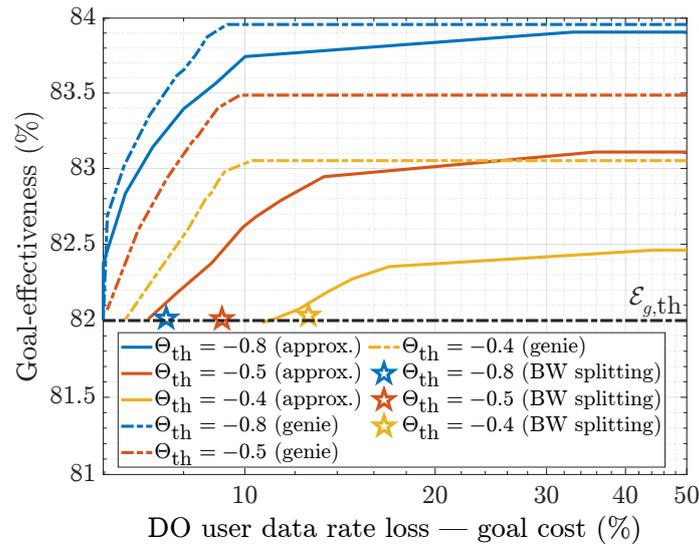
5. Numerical Evaluation

In this section, we evaluate the performance of the proposed GO optimization strategy in Algorithm 1 in the proposed coexistence network setting. To be able to compare the performance with fixed PER and P_d , we consider the same scenario of the previous evaluation (Figure 7). As a first result, in Figure 8, we show the trade-off between goal effectiveness and goal cost, for a fixed goal-effectiveness threshold $\mathcal{E}_g = 82\%$, two different values of total system bandwidth (Figure 8a with $W = 1$ GHz and Figure 8b with $W = 500$ MHz) and for different goal value thresholds (the same as the previous figure). Moreover, let us recall that our solution hinges on the problem approximation that replaces the first indicator function in (30) with its expectation computed on a validation set. Therefore, in Figure 8a,b, we compare this solution with the one involving a *genie* that uses a posteriori knowledge on the resulting entropy at the output of the classifier. Obviously, this solution cannot be implemented in practice, but it represent our benchmark, as it approaches the optimum of (19) as the trade-off parameter Ω increases. Finally, we compare our goal-oriented optimization method with a strategy that splits the bandwidth across the two systems, thus resulting in zero co-channel interference leaked from one system to another, however, at the cost of less radio bandwidth available for each of the two systems. In this case, only the PER of the GO user is selected dynamically. This can also be considered as a goal-oriented communication KPIs selection, but the interference to and from a DO system is not managed, as the system is assumed to not be affected by interference. Note that the bandwidth

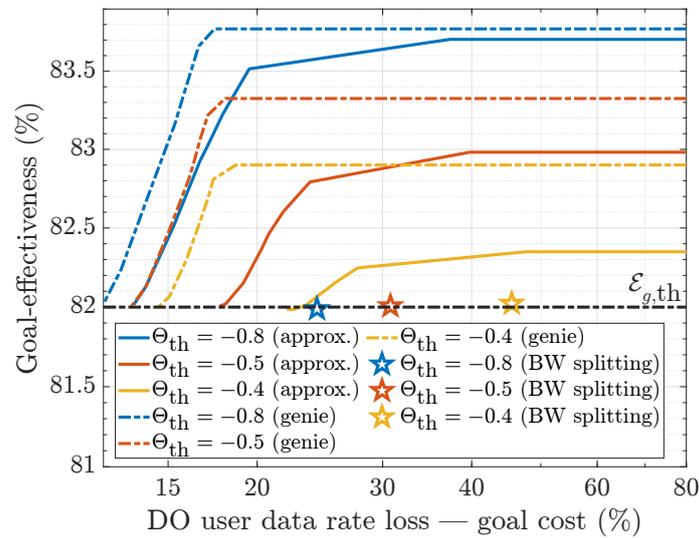
splitting is empirically selected to obtain the target level of goal effectiveness to compare the results in terms of goal cost. In other words, UE_g is allocated the minimum amount of bandwidth needed to achieve the target goal effectiveness, while the remaining portion of the bandwidth is allocated to UE_d . In the curves of Figure 8a, the hyper-parameter Ω increases from right to left. The curves with the same colors are obtained with the same system parameters, but they represent the three different strategies: (i) GO optimization (i.e., our proposed strategy) with solid lines, (ii) genie-aided GO optimization with dashed lines and (iii) bandwidth splitting scenario with the pentagrams. First, we can notice how our solution is able to always guarantee the goal-effectiveness constraint, with the goal cost decreasing towards the point in which the constraint is exactly attained (left-hand side of the plot). Moreover, as the goal value threshold (Θ_{th}) increases, the minimum achieved goal cost increases, as expected (see, e.g., yellow curve vs. blue curve). In particular, for the case with larger bandwidth (Figure 8a), the effect is more visible, but still limited, through the approximation, while the genie-aided solution is able to keep the cost close to optimal even when increasing Θ_{th} . However, as Θ_{th} decreases, the approximated solution becomes closer to the genie solution. Notably, *both non-orthogonal spectrum sharing* solutions outperform the one with bandwidth splitting, with considerable gain for the case of limited total available bandwidth (up to 50%—Figure 8a). This suggests that the power of a GO system optimization is even more relevant under communication resource scarcity, as also pointed out in the introduction of this work. Still, it should be noted the large difference between the GO communication KPIs selection and the legacy one, which is the purpose of the following performance evaluation.

As a second result, Figure 9 builds on the results of Figure 7 to compare our optimization method with the fixed PER and P_d setting (i.e., the non-optimized case). In particular, let us focus on Figure 9a, i.e., the case with $\Theta_{th} = -0.8$. We compare three strategies: (i) the fixed (i.e., non-adaptive) strategy of Figure 7 (black curve), (ii) a strategy with fixed PER, with our adaptive algorithm only controlling $P_{d,t}$ (red curve), and (iii) our adaptive strategy for both PER and P_d (blue pentagram). First, we can notice how strategy (i) (i.e., fixed PER and adaptive $P_{d,t}$) is shown only for a subset of possible PERs, i.e., only the feasible ones, given the target goal effectiveness. Then, from the red curve, we can notice how the proposed optimization method (although without adaptive PER control) is able to follow the profile of the non-optimized case, however slightly enlarging the feasibility region for some values of PER (e.g., $\gamma = 10^{-4}$), i.e., achieving lower goal costs, thanks to power control and adaptation. However, this gain is negligible in most of the cases. On the other hand, the full optimization strategy (i.e., the adaptive PER and P_d method), which is the main novelty of this work, achieves the lowest goal cost (around 6% vs. 12% of the fixed PER strategy), with a different resulting average PER. Similar considerations can be made for Figure 9b (with $\Theta_{th} = -0.5$), however with a much larger gain obtained by our strategy with respect to the two benchmarks. In addition, in Figure 9c, the two non-adaptive solutions do not provide feasible points for $\Theta_{th} = -0.4$, while our strategy is able to guarantee the target goal effectiveness with a goal cost around 8.5%, to be compared with the optimal value obtained by the genie (6.5%).

From Figure 9, we can conclude that it is highly beneficial to adapt the target PER (i.e., communication reliability of GO communication) to higher layer performance (i.e., application performance) rather than adapting communication parameters to keep, e.g., a fixed target PER. This can have a huge consequence in adaptive MCS mechanisms, whose aim would be to guarantee application performance (adapting PER and interference tolerance) rather than guaranteeing a target PER a priori, in the case of GO communications.



(a) Total system bandwidth $W = 1$ GHz



(b) Total system bandwidth $W = 500$ MHz

Figure 8. Trade-off between goal cost and goal effectiveness, with our method and with genie, against bandwidth splitting scenario.

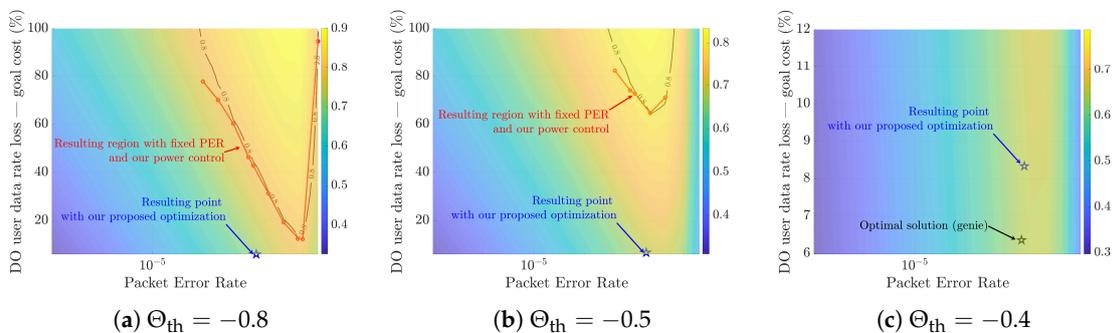


Figure 9. Goal-effectiveness heat maps as a function of goal cost (DO user data rate loss) and target PER of GO system, with and without optimization, for fixed E2E delay threshold ($D_{\max} = 45$ ms) and different goal value thresholds Θ_{th} .

Let us now focus on the adaptation capabilities of our proposed optimization. In particular, in Figure 10, we show a moving version of the goal effectiveness (i.e., the goal effectiveness estimated over the past 2000 slots—Figure 10a) and moving average goal cost (Figure 10b) as functions of time, introducing two unexpected events that need the algorithm to adapt to new conditions. Namely, at the beginning, the system requirements are $D_{\max} = 45$ ms, $\Theta_{\text{th}} = -0.4$ and $\mathcal{E}_g = 0.8$. Then, after 10,000 iterations, we switch from $\Theta_{\text{th}} = -0.4$ to $\Theta_{\text{th}} = -0.5$, keeping the same target E2E delay and goal effectiveness. After another 10,000 iterations (i.e., at slot 20,000), we switch from $\mathcal{E}_g = 0.8$ to $\mathcal{E}_g = 0.85$, while keeping the same goal value target and E2E delay requirement (i.e., $\Theta_{\text{th}} = -0.5$ and $D_{\max} = 45$ ms). First, from Figure 10a, we can notice how, from a goal-effectiveness perspective, the method is able to keep the desired value, switching to the new requirement after $t = 20000$. In addition, looking at Figure 10b, we can appreciate how the method always seeks the optimal solution in terms of goal cost. Indeed, when switching from $\Theta_{\text{th}} = -0.4$ to $\Theta_{\text{th}} = -0.5$, the goal cost reduces, thanks to the more relaxed goal value threshold. Then, when imposing a stricter goal-effectiveness constraint ($t = 20,000$), the goal cost increases again to attain the desired performance.

From Figure 10, it is then possible to appreciate the adaptation capabilities of the proposed method with respect to online requirement modifications. In particular, the method is able to autonomously detect this change through the virtual queue and take corresponding actions to stabilize the system towards desired performance.

We have shown how the method behaves in the presence of requirement changes, but how does it react to a non-stationary environment? Moreover, what is the effect of computation resource availability (i.e., computing delay) in the proposed scenario under investigation?

To answer these questions, we consider a system in which, after 10,000 iterations, due to an exogenous event, the computing time experiences an offset of +5 ms, i.e., it is systematically increased by 5 ms. Similarly, after 30,000 iterations, it is systematically increased by an additional 2 ms (i.e., 7 ms with respect to the starting point). The E2E delay constraint is set to $D_{\max} = 50$ ms. The results in terms of goal effectiveness and goal cost are shown in Figure 11a,b, respectively, with the same approach of Figure 10 (moving averages). From a goal-effectiveness perspective, we can notice how the method is able to attain the desired performance ($\mathcal{E}_g = 80\%$), despite a transient period after the non-stationarity appears in the system, highlighted by the dashed ellipses in the plot. At the same time, adapting to the new system conditions (higher computing time) increases the cost, as a lower communication delay is required to guarantee the delay constraint (we always assume that the inference input/ image upload rate is such that no classification request buffering at the MEH occurs, even when the delay of GO communication is substantially reduced). Indeed, the only way is to reduce interference, i.e., the DO system data rate. Again, the method is able to autonomously detect this change through the virtual queue and take corresponding actions to stabilize the system towards desired performance.

From Figure 11, we can conclude that *the computing resources available to a GO system have a strong impact on the cost of a DO system that coexists and interferes with the former*. This is a clear way of seeing the effect of connect-compute resource availability in goal-oriented communication systems for edge inference.

From the numerical results shown in this section, we can draw the following general conclusions:

- There exists a trade-off between goal effectiveness and goal value, with the latter being related to communication performance of a DO system coexisting with the GO system; our method is able to explore this trade-off, with close to optimal performance in different conditions, depending on the specific requirements (cf. Figure 8a).
- Fixing the PER (i.e., adapting communication to maintain an a priori target PER) while adapting the DO user transmit power does not provide much better performance than a strategy with both variables fixed across time (cf. Figure 9). Higher gains are achieved via a fully adaptive system.

- Our method is able to dramatically reduce the goal cost, while guaranteeing target goal effectiveness, by adaptively selecting target PER and DO user transmit power, based on measured application performance, even in the cases in which the fixed strategy fails to find a feasible solution (cf. Figure 9c).
- Changing requirements over time (e.g., because of new application constraints) does not prevent our method from adaptively allocating resources to attain new levels of goal effectiveness and/or goal values (cf. Figure 10). Moreover, the method works in both directions: it increases the cost when a transition to stricter requirements occurs, while it reduces the cost whenever requirements are relaxed. The latter, thanks to properly defined state variables (i.e., virtual queues), is able to capture the behavior of the system in terms of constraint violations. Obviously, this capability is limited to the cases in which non-stationarity occurs on a longer time scale than the method's adaptation.
- Computation resource availability at the GO system strongly affects the goal cost in terms of DO system data rate loss, a non-trivial result, never shown in the literature before, to the best of our knowledge. In addition, non-stationary environments, in terms of connect-compute resource availability, do not affect the adaptation capabilities of our method (cf. Figure 11).
- The proposed evaluations have been performed through system-level simulations and represent a first step towards a demonstration of a GO system optimization. The model-based analysis presented in this paper is useful to obtain insights on the potential gains that can be obtained in a real system. Obviously, a real-world demonstration would give rise to several new challenges, including the message passing between agents, translating to a coordination overhead that needs to be considered when designing an interface inter-connecting the systems. However, the obtained gains show the convenience of adopting such a GO approach for system optimization and lay the foundations for more practical works.

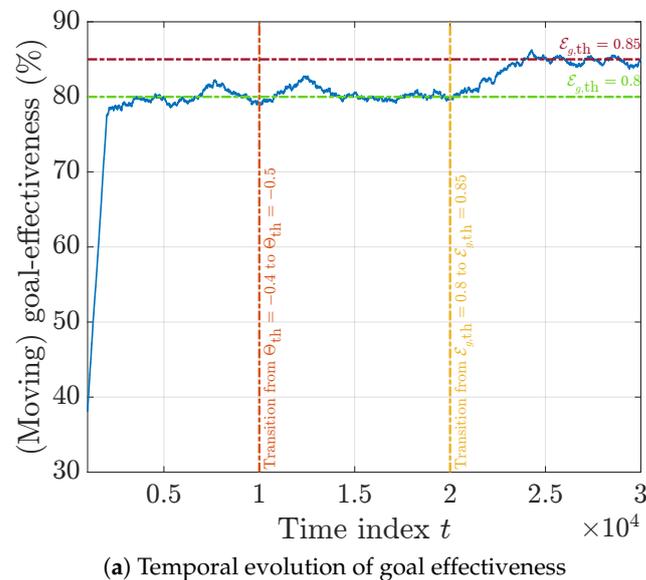


Figure 10. Cont.

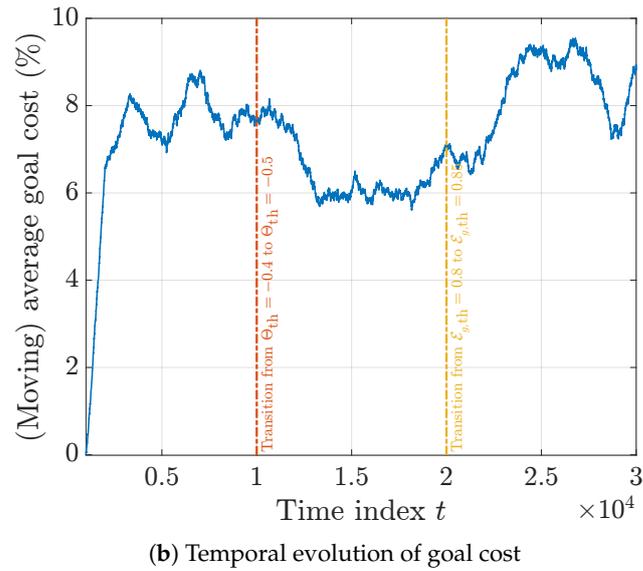


Figure 10. Adaptation capabilities of proposed method to online GO system requirement changes (i.e., Θ_{th} and ϵ_g).

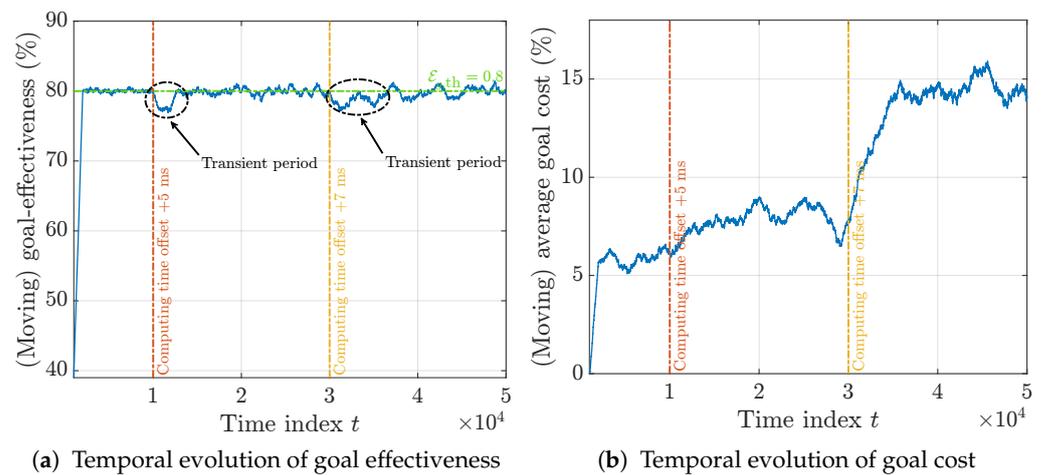


Figure 11. Adaptation capabilities of proposed method to compute resource availability (i.e., computing delay).

6. Conclusions and Future Directions

We have analyzed and optimized a wireless network scenario in which a GO and a legacy DO communication system coexist, fully sharing the same spectrum resources. While this has been proposed previously for the coexistence of DO and semantic communication systems, we focused on GO communication, also analyzing the effect of communication errors and availability of computing resources toward accomplishing the goal of communication.

We first explained the concept of GO communications, and we provided a general problem formulation approach to then tailor it to the proposed system scenario, showing how a GO resource allocation strategy can bring high gains in terms of overall system performance. The latter has been defined through two measures, goal effectiveness and goal cost, with the former translating into probability of confident inference on time, and the latter referring to the performance loss of the legacy system, in terms of data rate. The proposed GO approach is to adapt communication KPIs (i.e., PER) to the actual outcome of communication, measured at the application level. In addition to a first numerical performance evaluation for a non-optimized setting aimed at showing the potential trade-offs, our proposed optimization leverages application performance measures to update

suitably defined state variables, whose long-term stability has been exploited to achieve the goal with the lowest cost.

After proposing an algorithm with theoretical guarantees, we tested it on the proposed scenario, in which the two systems coexist and interfere with each other. It has been shown, through numerical simulations, the gain, in terms of DO system data rate for a given goal effectiveness of the GO system, of using an adaptive method for both communication reliability (i.e., PER) and legacy system user transmit power (i.e., affecting interference received by goal-oriented system and legacy system user data rate). Moreover, the proposed approach has been tested in non-stationary environments, showing good adaptation capabilities to new requirements and computing resource availability, showing the strong link between communication and computing in future networks. Overall, the paper shows the superiority of a GO selection of communication KPIs, based on communication and computation resource availability, as well as interference.

Future steps involve scenarios with multiple GO and legacy system users, the optimization of GO user transmit power and other communication related parameters, cooperative inference, but also other applications beyond edge inference. Finally, an interesting research direction is that of semantic and GO feature extraction into a unified framework in which only the most important features are transmitted and/or protected from wireless errors by the GO users. From a methodological point of view, the exploration of data-driven techniques such as DRL could help in solving the issue of not having a model relating communication reliability (e.g., PER) to the goal value. In this case, this relation would be learnt from experience, by the interaction of the agent(s) with the environment. Further, practical demonstrations of the proposed GO communication paradigm are worth being investigated, with clear challenges related to the validity of the models (including energy consumption, delay and inference confidence), the needed message exchange between agents to enable the proposed scheme and the interaction/interface between network and MEC operators, also from an economic perspective. Finally, we note that existing radio equipment, cybersecurity and privacy-related regulation and market access frameworks such as the Radio Equipment Directive [60], the EU Cybersecurity Act [61] and the upcoming Cyber Resilience Act [62] rely on classical communication approaches; new paradigms such as GO communications are considered insufficiently, and corresponding adaptations will be necessary.

Author Contributions: Conceptualization, all authors; methodology, M.M. and M.C.F.; software, M.M.; validation, all authors; writing—original draft preparation, M.M. and M.C.F.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956).

Data Availability Statement: The data presented in this study are available on request from the corresponding author (accurately indicate status).

Conflicts of Interest: Authors Miltiadis C. Filippou, Leonardo Gomes Baltar, and Markus Dominik Mueck were employed by the company Intel Deutschland GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Uusitalo, M.A.; Rugel, P.; Boldi, M.R.; Strinati, E.C.; Demestichas, P.; Ericson, M.; Fettweis, G.P.; Filippou, M.C.; Gati, A.; Hamon, M.H.; et al. 6G Vision, Value, Use Cases and Technologies From European 6G Flagship Project Hexa-X. *IEEE Access* **2021**, *9*, 160004–160020. [[CrossRef](#)]
2. Saad, W.; Bennis, M.; Chen, M. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Netw.* **2020**, *34*, 134–142. [[CrossRef](#)]
3. Hexa-X. *Deliverable D1.2—Expanded 6G Vision, Use Cases and Societal Values—Including Aspects of Sustainability, Security and Spectrum*; European Union: Luxembourg, 2021.
4. Hexa-X. *Deliverable D1.3—Targets and Requirements for 6G—Initial E2E Architecture*; European Union: Luxembourg, 2022.

5. Huo, Y.; Lin, X.; Di, B.; Zhang, H.; Hernando, F.J.L.; Tan, A.S.; Mumtaz, S.; Demir, Ö.T.; Chen-Hu, K. Technology Trends for Massive MIMO towards 6G. *Sensors* **2023**, *23*, 6062. [[CrossRef](#)] [[PubMed](#)]
6. Hexa-X. *Deliverable D4.2—AI-Driven Communication & Computation Co-Design: Initial Solutions*; European Union: Luxembourg, 2022.
7. Strinati, E.C.; Alexandropoulos, G.C.; Wymeersch, H.; Denis, B.; Sciancalepore, V.; D’Errico, R.; Clemente, A.; Phan-Huy, D.T.; De Carvalho, E.; Popovski, P. Reconfigurable, Intelligent, and Sustainable Wireless Environments for 6G Smart Connectivity. *IEEE Commun. Mag.* **2021**, *59*, 99–105. [[CrossRef](#)]
8. Sur, S.N.; Singh, A.K.; Kandar, D.; Silva, A.; Nguyen, N.D. Intelligent Reflecting Surface Assisted Localization: Opportunities and Challenges. *Electronics* **2022**, *11*, 1411. [[CrossRef](#)]
9. Ramezani, P.; Lyu, B.; Jamalipour, A. Toward RIS-Enhanced Integrated Terrestrial/Non-Terrestrial Connectivity in 6G. *IEEE Netw.* **2023**, *37*, 178–185. [[CrossRef](#)]
10. Kekki, S.; Featherstone, W.; Fang, Y.; Kuure, P.; Li, A.; Ranjan, A.; Purkayastha, D.; Jiangping, F.; Frydman, D.; Verin, G.; et al. ETSI White Paper: MEC in 5G networks. In *The European Telecommunications Standards Institute (ETSI)*; Tech. Rep. ETSI White Paper No. 28; ETSI: Sophia Antipolis, France, 2018.
11. Letaief, K.B.; Shi, Y.; Lu, J.; Lu, J. Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 5–36. [[CrossRef](#)]
12. Yang, H.; Alphones, A.; Xiong, Z.; Niyato, D.; Zhao, J.; Wu, K. Artificial-Intelligence-Enabled Intelligent 6G Networks. *IEEE Netw.* **2020**, *34*, 272–280. [[CrossRef](#)]
13. Zhang, S.; Zhu, D. Towards artificial intelligence enabled 6G: State of the art, challenges, and opportunities. *Comput. Netw.* **2020**, *183*, 107556. [[CrossRef](#)]
14. Park, J.; Samarakoon, S.; Bennis, M.; Debbah, M. Wireless network intelligence at the edge. *Proc. IEEE* **2019**, *107*, 2204–2239. [[CrossRef](#)]
15. Zhou, Z.; Chen, X.; Li, E.; Zeng, L.; Luo, K.; Zhang, J. Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. *Proc. IEEE* **2019**, *107*, 1738–1762. [[CrossRef](#)]
16. Calvanese Strinati, E.; Barbarossa, S. 6G networks: Beyond Shannon towards semantic and goal-oriented communications. *Comput. Netw.* **2021**, *190*, 107930. [[CrossRef](#)]
17. Goldreich, O.; Juba, B.; Sudan, M. A Theory of Goal-Oriented Communication. *J. ACM* **2012**, *59*, 1–65. [[CrossRef](#)]
18. Zou, H.; Zhang, C.; Lasaulce, S.; Saludjian, L.; Poor, V. Goal-Oriented Quantization: Analysis, Design, and Application to Resource Allocation. *arXiv* **2022**, arXiv:2209.15347.
19. Binucci, F.; Banelli, P.; Di Lorenzo, P.; Barbarossa, S. Dynamic Resource Allocation for Multi-User Goal-oriented Communications at the Wireless Edge. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 20 August–2 September 2022; pp. 697–701.
20. Zhang, C.; Zou, H.; Lasaulce, S.; Saad, W.; Kountouris, M.; Bennis, M. Goal-Oriented Communications for the IoT and Application to Data Compression. *arXiv* **2022**, arXiv:2211.05378.
21. Pezone, F.; Barbarossa, S.; Di Lorenzo, P. Goal-Oriented Communication for Edge Learning Based On the Information Bottleneck. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8832–8836.
22. Pappas, N.; Kountouris, M. Goal-Oriented Communication For Real-Time Tracking In Autonomous Systems. In Proceedings of the 2021 IEEE International Conference on Autonomous Systems (ICAS), Montreal, QC, Canada, 11–13 August 2021; pp. 1–5.
23. Mostaani, A.; Simeone, O.; Chatzinotas, S.; Ottersten, B. Learning-based Physical Layer Communications for Multiagent Collaboration. In Proceedings of the 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 8–1 September 2019; pp. 1–6. [[CrossRef](#)]
24. Merluzzi, M.; Martino, A.; Costanzo, F.; Di Lorenzo, P.; Barbarossa, S. Dynamic Ensemble Inference at the Edge. In Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 7–11 December 2021; pp. 1–6. [[CrossRef](#)]
25. Merluzzi, M.; Battiloro, C.; Di Lorenzo, P.; Strinati, E.C. Energy-Efficient Classification at the Wireless Edge with Reliability Guarantees. In Proceedings of the 2022 IEEE International Conference on Communications Workshops (ICC Workshops), Seoul, Republic of Korea, 16–20 May 2022; pp. 109–114. [[CrossRef](#)]
26. Merluzzi, M.; Di Lorenzo, P.; Barbarossa, S. Wireless Edge Machine Learning: Resource Allocation and Trade-Offs. *IEEE Access* **2021**, *9*, 45377–45398. [[CrossRef](#)]
27. Shao, J.; Zhang, J. BottleNet++: An End-to-End Approach for Feature Compression in Device-Edge Co-Inference Systems. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 1–7 June 2020; pp. 1–6. [[CrossRef](#)]
28. Xie, H.; Qin, Z.; Tao, X.; Letaief, K.B. Task-Oriented Multi-User Semantic Communications. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 2584–2597. [[CrossRef](#)]
29. Sana, M.; Strinati, E.C. Learning Semantics: An Opportunity for Effective 6G Communications. In Proceedings of the 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 8–11 January 2022; pp. 631–636. [[CrossRef](#)]
30. Lee, C.H.; Lin, J.W.; Chen, P.H.; Chang, Y.C. Deep Learning-Constructed Joint Transmission-Recognition for Internet of Things. *IEEE Access* **2019**, *7*, 76547–76561. [[CrossRef](#)]

31. Merluzzi, M.; Filippou, M.C.; Baltar, L.G.; Strinati, E.C. Effective Goal-oriented 6G Communications: The Energy-aware Edge Inferencing Case. In Proceedings of the 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Grenoble, France, 7–10 June 2022; pp. 457–462. [CrossRef]
32. Mu, X.; Liu, Y. Exploiting Semantic Communication for Non-Orthogonal Multiple Access. *arXiv* **2022**, arXiv:2209.06006.
33. Maman, M.; Calvanese-Strinati, E.; Dinh, L.N.; Haustein, T.; Keusgen, W.; Wittig, S.; Schmieder, M.; Barbarossa, S.; Merluzzi, M.; Costanzo, F.; et al. Beyond private 5G networks: Applications, architectures, operator models and technological enablers. *EURASIP J. Wirel. Commun. Netw.* **2021**, *2021*, 195. [CrossRef]
34. Lee, H.; Ahn, H.; Park, Y.D. Performance analysis of coexistence of traditional communication system and emerging semantic communication system. *ICT Express* **2022**, *9*, 420–426. [CrossRef]
35. Kountouris, M.; Pappas, N. Semantics-Empowered Communication for Networked Intelligent Systems. *IEEE Commun. Mag.* **2021**, *59*, 96–102. [CrossRef]
36. Mu, X.; Liu, Y. Semi-NOMA enabled Coexisting Semantic and Bit Communications. In Proceedings of the 2022 International Symposium on Wireless Communication Systems (ISWCS), Hangzhou, China, 19–22 October 2022; pp. 1–6. [CrossRef]
37. Hollnagel, E. A Goals-Means Task Analysis Method. 1991. Available online: <https://erikhollnagel.com/onewebmedia/GMTA.pdf> (accessed on 20 July 2023).
38. Popovski, P. Ultra-reliable communication in 5G wireless systems. In Proceedings of the 1st International Conference on 5G for Ubiquitous Connectivity, Levi, Finland, 26–27 Nov. 2014 ; pp. 146–151. [CrossRef]
39. Merluzzi, M.; Pietro, N.d.; Di Lorenzo, P.; Strinati, E.C.; Barbarossa, S. Discontinuous Computation Offloading for Energy-Efficient Mobile Edge Computing. *IEEE Trans. Green Commun. Netw.* **2022**, *6*, 1242–1257. [CrossRef]
40. Battiloro, C.; Lorenzo, P.D.; Merluzzi, M.; Barbarossa, S. Lyapunov-based Optimization of Edge Resources for Energy-Efficient Adaptive Federated Learning. *IEEE Trans. Green Commun. Netw.* **2022**, *7*. [CrossRef]
41. Sana, M.; De Domenico, A.; Yu, W.; Lostanlen, Y.; Calvanese Strinati, E. Multi-Agent Reinforcement Learning for Adaptive User Association in Dynamic mmWave Networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6520–6534. [CrossRef]
42. Yu, H.; Neely, M.J. Dynamic Transmit Covariance Design in MIMO Fading Systems With Unknown Channel Distributions and Inaccurate Channel State Information. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 3996–4008. [CrossRef]
43. Mao, Y.; Zhang, J.; Letaief, K.B. Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 3590–3605. [CrossRef]
44. Mao, Y.; Zhang, J.; Song, S.H.; Letaief, K.B. Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems. *IEEE Trans. Wireless Commun.* **2017**, *16*, 5994–6009. [CrossRef]
45. Merluzzi, M.; Di Lorenzo, P.; Barbarossa, S.; Frascolla, V. Dynamic Computation Offloading in Multi-Access Edge Computing via Ultra-Reliable and Low-Latency Communications. *IEEE Trans. Signal Inf. Process. Over Netw.* **2020**, *6*. [CrossRef]
46. Shlezinger, N.; Farhan, E.; Morgenstern, H.; Eldar, Y.C. Collaborative Inference via Ensembles on the Edge. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 8478–8482. [CrossRef]
47. Mitola, J.; Maguire, G.Q. Cognitive radio: Making software radios more personal. *IEEE Pers. Commun.* **1999**, *6*, 13–18. [CrossRef]
48. Haykin, S. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.* **2005**, *23*, 201–220. [CrossRef]
49. Luo, J.; Zhao, S.; Yang, Z.; Huang, G.; Tang, D. Data-Driven-Based Relay Selection and Cooperative Beamforming for Non-Regenerative Multi-Antenna Relay Networks. *IEEE Access* **2021**, *9*, 167154–167161. [CrossRef]
50. Ibrahim, L.N.; Al-Mistarihi, M.F.; Khodeir, M.A.; Alhulayil, M.; Darabkh, K.A. Best Relay Selection Strategy in Cooperative Spectrum Sharing Framework with Mobile-Based End User. *Appl. Sci.* **2023**, *13*, 8127. [CrossRef]
51. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
52. Sun, C.; She, C.; Yang, C.; Quek, T.Q.S.; Li, Y.; Vucetic, B. Optimizing Resource Allocation in the Short Blocklength Regime for Ultra-Reliable and Low-Latency Communications. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 402–415. [CrossRef]
53. She, C.; Yang, C.; Quek, T.Q.S. Radio Resource Management for Ultra-Reliable and Low-Latency Communications. *IEEE Commun. Mag.* **2017**, *55*, 72–78. [CrossRef]
54. Canziani, A.; Paszke, E.C. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv* **2016**, arXiv:1605.07678.
55. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
56. Saerens, M.; Latinne, P.; Decaestecker, C. Any reasonable cost function can be used for a posteriori probability approximation. *IEEE Trans. Neural Netw.* **2002**, *13*, 1204–1210. [CrossRef]
57. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical Report 0; University of Toronto: Toronto, ON, Canada, 2009.
58. Neely, M.J. *Stochastic Network Optimization with Application to Communication and Queueing Systems*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2010.
59. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
60. European Union Directive 2014/53/EU OF THE European Parliament and of the Council of 16 April 2014 on the Harmonisation of the Laws of the Member States Relating to the Making Available on the Market of Radio Equipment and Repealing Directive 1999/5/EC. 2014. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0053> (accessed on 20 July 2023).

61. Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on Information and Communications Technology Cybersecurity Certification and Repealing Regulation (EU) No 526/2013 (Cybersecurity Act). 2019. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0881> (accessed on 20 July 2023).
62. Proposal for a Regulation OF THE European Parliament and of the Council on Horizontal Cybersecurity Requirements for Products with Digital Elements and Amending Regulation (EU) 2019/1020. 2022. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0454> (accessed on 20 July 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.