Analog versus Digital Pulse Amplitude Modulation for Goal-Oriented Wireless Communications

Francesco Binucci[1], Paolo Banelli[1], Paolo Di Lorenzo[2], Sergio Barbarossa[2]
[1]Department of Engineering, University of Perugia, Via G. Duranti 93, 06128, Perugia, Italy
[2]DIET department, Sapienza University of Rome, via Eudossiana 18, 00184, Rome, Italy

***Abstract*—Goal-Oriented communications are an emerging paradigm for wireless edge intelligence applications. Within this framework, this paper compares analog versus digital modulations for over-the-air classification tasks, under optimal resource management policies that trade energy consumption, delay, and accuracy. The resource allocation problem is formulated using Lyapunov stochastic optimization, and is solved in an online fashion with limited complexity. Simulation results illustrate superior performance of analog designs when the task arrival process pushes digital communications closer to channel capacity.**

***Index Terms*—Goal-oriented communications, resource allocation, stochastic optimization, pulse amplitude modulation.**

I. Introduction

The new generation of mobile networks will support a plethora of services, mainly based on Artificial Intelligence (AI) and Machine Learning (ML), which typically need low latency and high reliability, consequently demanding for high computational and energetic resources. These requirements are often in contrast with the limited capabilities of the User Equipments (UEs) (e.g. cameras, sensors, etc.) [1] [2]. To overcome these issues, Edge Intelligence (EI) moves computational resources to Edge Servers (ESs) that are placed closer to the UEs, thus enabling low-latency connect-compute services such as computation offloading. The exponential growth of data-traffic exchanged to enable the aforementioned AI services challenge wireless communication systems to make smart use of transmission resources, possibly adapting the communication paradigm to these new needs. In this context, *Goal-Oriented* Communications (GOCs) and *Semantic-Oriented* communications represent emerging hot-topics [1], [3], [4] that mitigate the unsustainable increase of communication resources by transmitting the minimum information that is necessary to perform a specific task, while guaranteeing a target accuracy and reliability. For instance, [5] suggests a holistic system view, where communication, computation, learning, and control are jointly managed in order to optimize the latency, the reliability, and the system energy consumption.

Related Works. Seminal works on EI explored the trade-off between energy, latency and learning accuracy [5]–[7]. Recently, a GOC framework exploiting the Gaussian Information Bottleneck (GIB) [8], [9] was proposed in [10] for regression tasks. To overcome the limited applicability of GIB to classification tasks, and non Gaussian statistics, [11], [12] proposed an IB-inspired *digital* GOC framework for image classification based on Convolutional Encoders (CEs) coupled with proper Convolutional Classifiers (CCs) at the ES, which can optimally manage resources also in BER-impaired systems [13]. Many other GOCs systems has been proposed in the recent literature, such as [4] that considers a variational IB principle. Another possibility is to couple GOC with Deep Joint Source Channel Coding (DJSCC) [14], [15] where, rather than employing the classical Shannon separation theorem, source and channel coding are jointly performed using specific neural networks, whose (*analog*) outputs are directly mapped to the transmitted signal, showing superior performance of analog solutions in low signal-to-noise ratio (SNR) regimes. New and more practically implementable approaches have been recently proposed in [16]–[18], where DJSCC is deployed within digital communication systems.

Our contribution. Aim of this work is to investigate benefits of analog modulations in resource management and optimization of GOC systems. Thus, differently from DJSCC approaches in [14], [15], [17], and references therein, we focus on the optimal management of *computation and transmission* resources for *analog*-PAM GOC systems, with the aim of striking the best trade-off between energy, latency, and accuracy. We formulated a long-term optimization problem that is solved in an online fashion using Lyapunov stochastic optimization tools, without requiring any a priori statistical knowledge of channels and data arrivals. An (*ideal*) digital-PAM counterpart, e.g., (zero-BER) pulse code modulation (PCM), has been used for a fair and preliminary comparison with digital-GOCs. Numerical simulations illustrate the advantages of using direct analog-PAM GOCs with respect to digital ones, especially at low SNRs.

II. System Model

We consider a single UE that offloads a classification task to a single ES, as illustrated in Fig. 1. As proposed in [11]–[13], the UE's Data Units (DUs) X to be classified, are firstly compressed into the features W exploiting one among a set of convolutional encoders (CEs), each one characterized by a specific compression $|W|/|X| = 1/\rho^2$, where $|\cdot|$ is the size operator, and ρ is the compression factor in each image dimension. This way, the system is capable to retrieve

This work was supported by MIUR under the PRIN Liquid Edge Contract. The work of Paolo Di Lorenzo and Sergio Barbarossa was supported by the EU H2020 RISE-6G project under the grant number 101017011

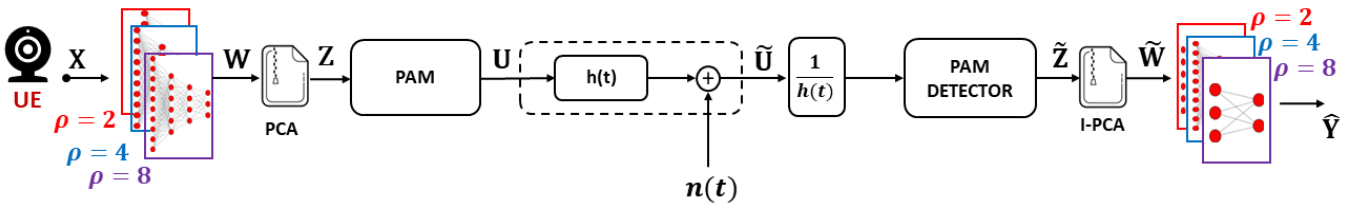

Fig. 1. Scheme of the proposed analog goal-oriented communication paradigm.

a first compact representation W that, according to a GOC philosophy, has to be as much informative as possible with respect to the inference task Y. Then, to further reduce the load of the wireless communication link, [11]–[13] employed another (intrinsically digital) compression codec, i.e., JPEG, whose output bits are finally mapped on the finite alphabets (e.g., M-QAM) of a digital GOC system. Differently, to possibly avoid the digital compression stage and analog to digital (A/D) conversion, herein we substitute JPEG with a Principal Component Analysis (PCA) projection. The PCA output Z directly generates the analog symbols, which are wirelessly transmitted by analog Pulse Amplitude Modulation (PAM) [19], and finally passed to the remote ES for classification. As in [11]–[13], the ES is equipped with a bank of CCs, each one associated by the compression factor ρ to a CE at the UE[1]. Each CE-CC couple is jointly trained offline to maximize the classification accuracy, as described in the sequel.

A. Analog transmission model

The base-band PAM signal is expressed by [19]

$$u(\tau) = \sum_{k=-\infty}^{+\infty} a_k p_g(\tau - kT_s), \tag{1}$$

where $a_k \in \mathbb{C}$ is the k-th transmitted symbol, $p_g(\tau)$ is the pulse shaper, and $S_r = 1/T_s$ is the symbol rate. We restrict our analysis to Inter-Symbol Interference (ISI)-free systems, by employing a Squared Root Raised Cosine (SRRC) shaping pulse $p_g(\tau)$, with two-sided baseband noise-equivalent bandwidth $B = S_r$, roll-off factor $\beta = 0.25$ [19], and unit energy. We consider analog modulations, where the PCA-compressed scalar outputs z_k are multiplexed on the I-Q components, and *directly* mapped to the PAM symbols by $a_k = z_{2k-1} + jz_{2k}$. Differently, for digital-PAM we want to compare with, the PCA outputs should first pass through n_b-bits A/D converters, before being mapped to finite digital modulation alphabets a_k (e.g., M-QAM). This way, for each image X the analog system is loaded by K symbols, with an overall compression ratio $C_r = K/|X|$, while the digital-PAM system is loaded by Kn_q bits, which corresponds to $Kn_q/\log_2(M)$ M-QAM symbols of the digital-PAM. In both scenarios, the designer could try to (dynamically) optimize the number of PCA features K (and number n_q of quantization bits in digital-PAM), for any possible CE-CC classification couple, i.e., compression factor ρ. Although goal-oriented quantization is a timely and meaningful design goal [20], for simplicity and space limitation herein we exploit the results of the digital system in [11], where CE outputs are efficiently converted by JPEG to a (compressed) bit representation, which has shown to be robust in a goal-oriented sense, e.g., accuracy-wise. Specifically, we observed there that the overall CE-JPEG compression ratio $C_r = K(\rho)/(|W|\rho^2)$ was scaling almost linearly with $1/\rho$. Thus, herein we fix the features $K(\rho)$ for each compression factor ρ, such that $C_r = \lfloor 1.6\%/\rho \rfloor$, which grants classification accuracy in $[60-94]\%$, for the chosen data-set. This way, considering RGB images with size $|X| = 256 \times 256 \times 3$, the PCA features are $K \in \{1573, 786, 393, 197, 98, 49\}$ for $\rho \in \{2, 4, 8, 16, 32, 64\}$, respectively. The PAM transmitted signal $u(\tau)$ is received through a noisy flat-fading wireless channel $h_c(\tau)$ and, after proper SRRC matched-filtering and one-tap equalization, it is passed to the ES for classification. We assume the system evolves in a time-slotted fashion, where each time-slot t has a fixed duration T. Specifically, in each time-slot, we assume a constant channel attenuation $h(t)$, and we aim to deploy resource management policies based on the instantaneous SNR

$$\gamma(t) = \frac{p_{tx}(t)|h(t)|^2}{B(t)N_0}, \tag{2}$$

where N_0 is the noise PSD, $B(t)$ is bandwidth, and $p_{tx}(t) = E\{|a_k|^2\}P_{avg}$ is the TX power, with $P_{avg} = \frac{1}{N}\sum_{n=1}^{N_t} \frac{||Z_n||^2}{K(\rho)}$, denoting the average power of the PCA compressed training set $\{Z_n\}_{n=1}^{N_t}$. In each slot t, we aim to optimize the transmit power, the bandwidth, and the compression factor ρ.

B. Training Procedure

First, each pair of CE-CCs is trained offline by minimizing $\frac{1}{N_t}\sum_{n=1}^{N_t} \mathcal{L}_{ce}(Y_n, \widehat{Y_n}; \phi, \theta)$, with respect to the CE's and CC's parameters θ and ϕ, respectively, where $\mathcal{L}_{ce}$ denotes the cross-entropy loss. Then, we train the PCA compression with the outputs of the pre-trained CC. Finally, we freeze the CEs and PCA, and we re-train the ES's CCs, adding to the PCA output a white Gaussian noise, to mimic different SNR conditions.

C. Latency Model

To control our dynamic system and quantify the overall delay experienced by a DU before processing, we introduce a *compression* and *transmission* queue at the UE side and a *classification* queue at the ES side. Furthermore, we assume that:

[1]The CEs and the CCs exploit multiple stages of convolutional and 2×2 max-pooling layers. The convolutional part of the CCs is followed by two fully connected layers, with a *softmax()* at the output. All the CEs have less than 10^3 parameters while the CCs have from 3×10^3 to 6×10^5 parameters.

i) a DU compressed at the t-th time-slot must be transmitted during the same time-slot. ii) the UE can transmit already compressed DUs while another one is being compressed.

The numbers $N_{tx}(t)$ and $N_c(t)$ of DUs that can be transmitted and compressed, respectively, in the t-th time-slot are

$$N_{tx}(t) = \left\lfloor \frac{2B(t)T}{K(\rho(t))(1+\beta)} \right\rfloor, \quad N_c(t) = \lfloor T f_d(t) J_d(\rho(t)) \rfloor, \tag{3}$$

where $2B(t)/(1+\beta)$ are the (real) features transmitted per second during the t-th time-slot, f_d is the UE-clock frequency, and $J_d(\rho)$ specifies the number of clock cycles that are necessary to compress a DU (which are stored in a LUT indexed by the compression factor ρ, and can be either evaluated experimentally or computed algorithmically). Recalling now assumptions i) and ii), and the impossibility to simultaneously compress & transmit the first DU, the number of processable DUs at the UE is expressed by

$$N_{UE}(t) = \left\lfloor \frac{T - 1/f_d(t)J_d(\rho(t))}{[K(\rho(t))(1+\beta)]/2B(t)} \right\rfloor. \tag{4}$$

Thus, the UE queue evolves according to

$$Q_{UE}(t+1) = \max(0, Q_{UE}(t) - N_{UE}(t)) + A(t), \tag{5}$$

where $A(t)$ represents the arrival process of new DUs.

By denoting $\frac{1}{J_s(\rho)}$ the clock-cycles needed to classify a DU compressed with specific ρ, the maximum number $N_{ES}(t)$ of DUs, which the ES can process during the t-th time-slot, is obtained by taking the maximum number of the (oldest) DUs stored in the ES such that the total clock-cycles $\sum_{i=1}^{N_{ES}(t)} \frac{1}{J_s(P(i))}$, to classify all of them, is lower than the maximum number $T f_c(t)$ of clocks-cycles the ES can perform. Then, the ES queue evolves according to

$$\begin{aligned} Q_{ES}(t+1) = &\max(0, Q_{ES}(t) - N_{ES}(t)) \\ &+ \min(Q_{UE}(t), N_{UE}(t)). \end{aligned} \tag{6}$$

We set $Q_{tot}(t) = Q_{UE}(t) + Q_{ES}(t)$ to quantify the overall delay. Indeed, by Little's law, for a fixed DU's arrival rate $\overline{A}$, the average delay can be written as $\overline{D}_{tot} = \frac{\mathbb{E}\{Q_{tot}\}}{\overline{A}}$.

D. Energy Model

The PAM-based system employs a transmission energy $E_{tx}(t) = T p_{tx}(t)$, while the computational energies at the UE and ES are modelled by [21],

$$E_c(t) = T\kappa_d f_d(t)^3, \quad E_s(t) = T\kappa_s f_c(t)^3, \tag{7}$$

where κ represents the *effective switched capacitance*. Thus, the overall system energy is summarized by

$$E_\alpha(t) = \alpha(E_{tx}(t) + E_c(t)) + (1-\alpha)E_s(t), \tag{8}$$

where $\alpha \in [0,1]$, allows to implement a UE-centric energy optimization, when $\alpha \to 1$, or an ES-centric one, when $\alpha \to 0$.

E. Accuracy Model

Similarly to [11]–[13], we model the classification accuracy by a LUT $G(\gamma(t), \rho(t))$, which is estimated offline on the test-set, for a set of different SNRs γ and compression factors ρ.

III. Problem formulation and solution

The aim of the proposed resource allocation strategy is to minimize, on the average, the overall energy consumption of the system under average accuracy and latency constraints. According to the Lyapunov framework [22], we start by formulating a long-term energy optimization problem

$$\begin{aligned} \min_{\Phi(t)} \quad & \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{E_\alpha(t)\} \\ \text{s.t.} \quad & (a)\ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{Q_{tot}(t)\} \le Q_{avg} \\ & (b)\ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{G(\rho(t), \gamma(t)\} \ge G_{avg} \\ & (d)\ p_{tx}(t) = \frac{\gamma(t) N_0 B(t)}{|h(t)|^2} \le p_{max}, \quad 0 \le B(t) \le B_{max} \\ & (e)\ \rho(t) \in \mathcal{S}, \quad f_s(t) \in \mathcal{F}_s, \quad f_d(t) \in \mathcal{F}_d, \quad \gamma(t) \in \mathbb{R}_0^+ \end{aligned} \tag{9}$$

where $\Phi(t) = [B(t), p_{tx}(t), f_d(t), \rho(t), \gamma(t), f_c(t)]$ is the set of the optimization variables. $(a), (b)$ are the *long-term* latency and accuracy constraints, respectively; (d) encompasses the constraints on the transmission power and bandwidth, while (e) represents all the feasibility constraints for the discrete optimization variables. We denote by $\mathcal{S}$ the set of the employable compression factors, while $\mathcal{F}_s$ and $\mathcal{F}_d$ are the sets of the clock frequencies of the ES and the UE respectively.

According to Lyapunov Optimization, we can derive an instantaneous optimization procedure that, on the basis of the system conditions (queue status, channel gain, etc.), performs a resource allocation policy that asymptotically tends to the optimal solution of (9) [22]. The long-term constraints $(a), (b)$, are handled by *virtual* queues defined as

$$\begin{aligned} Z(t+1) &= \max(0, Z(t) + \mu(Q_{tot}(t) - Q_{avg})) \\ Y(t+1) &= \max(0, Y(t) + \nu(G_{avg} - G(t)), \end{aligned} \tag{10}$$

where μ, ν are step-sizes used to improve the convergence speed of the algorithm. Then we build the Lyapunov function and the associated Lyapunov Drift plus penalty function

$$\begin{aligned} L(t) &= Y(t)^2 + Z(t)^2 \\ \Delta_p(t) &= \Delta(t) + V\mathbb{E}\{E_{tot}(t) | \Theta(t)\}, \end{aligned} \tag{11}$$

where $\Theta(t) = \{Z(t), Y(t)\}$, $\Delta(t) = \mathbb{E}\{L(t+1) - L(t)\}$ is the Lyapunov Drift function [22], and the parameter V controls the trade-off between objective function minimization and long-term constraints satisfaction.

Removing expectations and exploiting upper-bounds [22], omitted due the lack of space, the optimization decouples between UE and the ES. Specifically, the UE's optimization problem is

$$\begin{aligned} \min_{\Phi_d} \quad & -\frac{2TQ_{TX}B}{(1+\beta)K(\rho)} + TV\alpha\left(\gamma\frac{N_0 B}{|h|^2} + \kappa f_d^3\right) - \nu Y G(\gamma, \rho) \\ \text{s.t.} \quad & (a)\ 0 \le B \le B'_{max} \\ & (b)\ \rho \in \mathcal{S}, \quad f_d \in \mathcal{F}_d, \quad \gamma \in \Gamma \subset \mathbb{R}_0^+, \end{aligned} \tag{12}$$

TABLE I
CHANNEL SETTING.

$D_{max}[m]$	$f_c[GHz]$	$B_{max}[kHz]$	$E\{\|h\|^2\}$ [dB]
200	7.5	200	130

where $\Phi_d = [f_d, \gamma, \rho, B, p_{tx}]$ is the set of the UE optimization variables and $Q_{TX} = 2\mu^2(Q_{UE} - Q_{ES}) + \mu Z$. The left-hand side of constraint (d) in (9) has been directly substituted in the objective function, while the right-hand side of the same constraint has been captured in $B'_{max} = min\left(B_{max}, \frac{Q_{UE}K(\rho)(1+\beta)}{2T}, p_{max}\frac{\gamma|h|^2}{N_0 B}\right)$, which is the minimum between the maximum available bandwidth at the UE side, the bandwidth necessary to empty the transmission queue, and that one that satisfies constraint (d).

Problem (12) is mixed-integer, since the variables (f_d, ρ) lie in discrete sets. Furthermore, for any fixed couple (f_d, ρ), (12) is still non-convex, due to γB in the objective function. However, the simulation results in Fig. 2 suggest that the accuracy curves maybe safely fitted by concave functions with respect to the SNR γ: this way, problem (12) would be separately convex with respect to B and γ, and it could be solved by iterative alternating optimization procedures. However, since we evaluated the classification performance on a limited (discrete) set Γ of SNRs, as shown in Fig. 2, we threat also γ as a discrete optimization variable, thus reducing the problem to a linear program with respect to the transmission bandwidth B. Specifically, for any fixed triple (f_d, γ, ρ) the objective function becomes

$$f(B) = \left[V\frac{\gamma N_0}{|h|^2} - \frac{2Q_{TX}}{(1+\beta)K(\rho)}\right] TB - \lambda(f_d, \gamma, \rho). \quad (13)$$

When the ratio $m = \frac{\gamma N_0}{|h|^2} \leq \frac{2Q_{TX}}{V(1+\beta)K(\rho)}$, the objective function is linear and decreasing, with optimal solution given by $(p^*_{tx}, B^*) = (mB'_{max}, B'_{max})$. Otherwise, the objective function is increasing, and the optimal solution is $(p^*_{tx}, B^*) = (0, 0)$, which corresponds to avoid the transmission. Thus, the global optimal solution is obtained by computing the closed form solution for each couple (γ, ρ). Then, for any fixed compression ρ, the UE's clock-frequency f_d is simply the minimum one that grants $N_c(t) \geq N_{tx}(t)$. Finally, the solution is obtained by the triple (γ, ρ, f_d) that minimizes (13).

The decoupled ES's optimization problem turns out to be

$$\min_{f_c \in \mathcal{F}_s} \quad -Q_C N_{ES} + TV(1-\alpha)\kappa f_c^3, \quad (14)$$

where $Q_C = 2\mu^2 Q_{ES} + \mu Z$, and is solved by a short exhaustive search for $f_c \in \mathcal{F}_s$.

IV. SIMULATION RESULTS

We considered a flat fading channel characterized by a Jackes-Clarke auto-correlation function [23]. The fading model is assumed to be Rayleigh with zero mean and unit variance. The average path loss has been set according to the *ABG* model in [24], as summarized in Tab.I. A key novelty of this work resides in the UE's dynamic resources allocation policy, whose effectiveness can be better highlighted considering a pure UE-centric optimization strategy (i.e., $\alpha = 1$),

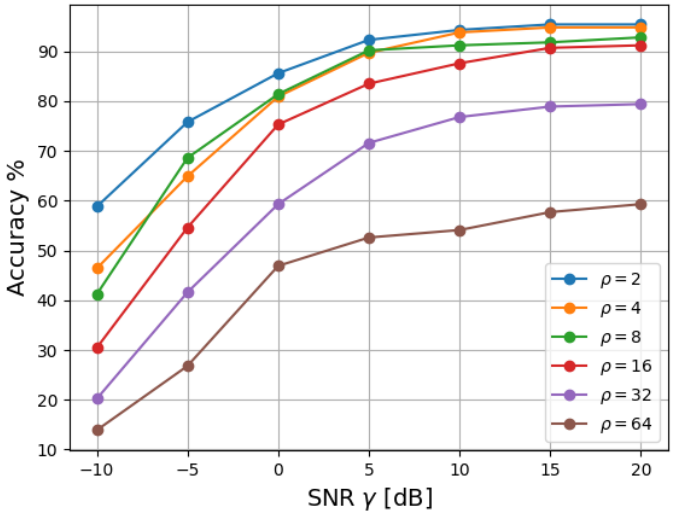

Fig. 2. Accuracy vs SNR for different ρ.

as we did in the simulations. We assumed $\kappa_{dev} = \kappa_{ser} = 1.097 \times 10^{-27}$ $[\frac{s}{cycles}]^3$, $\mathcal{F}_d = [0.1, 0.2, \ldots, 1] \times 1.4$ GHz and $\mathcal{F}_s = [0.1, 0.2, \ldots, 1] \times 4.5$ GHz. We set the maximum TX power to $p_{max} = 1$ W, and the time-slot duration to $T = 50$ ms, which fits the channel coherence time.

A. Performance of analog goal-oriented compression

We trained our GOC framework on a subset of the GTSRB data-set [25], with 43 classes, composed of 1213 RGB images of traffic signs, divided in 776 images for training, 194 for validation, and 243 for the test. The CE-CC classifiers have been trained for the analog-PAM system with the SNRs $\gamma \in [-10, 20]$ dB shown in Fig. 2, as described in sec. II-B. Fig.2 shows a graceful accuracy degradation, which is acceptable also at rather low SNRs for several compression factors ρ. This behaviour, already highlighted in [14], [15], represents one of the most important strength of the (analog) direct modulation schemes, since digital M-QAM systems request higher SNRs to avoid BER, which could excessively deteriorate classification accuracy.

B. Performance comparison

We tested the proposed analog dynamic resource allocation strategy for different accuracy constraints and different image arrivals rates, under a delay constraint equal to 200 ms. We compared the results with the digital M-QAM counterpart described in Section II, whose resource optimization slightly modifies [11], and is not detailed due to lack of space. Specifically, we considered an ideal, capacity achieving, M-QAM PAM with zero-BER, where the bandwidth and the modulation index are jointly optimized, under the same constraints. The M-QAM system employs $n_q = 8$ quantization bits, which turned out to be the best choice in this scenario. To make sense of the results, which highly depend on the channel conditions and the design parameters, we define the loading factor metrics

$$\eta_{dig}(t) = \frac{n_q A(t) K(\rho(t))/\tau}{C_{max}(t)}, \quad \eta_{ana}(t) = \frac{A(t)/\tau(1+\beta)}{2B_{max}/K(\rho(t))},$$

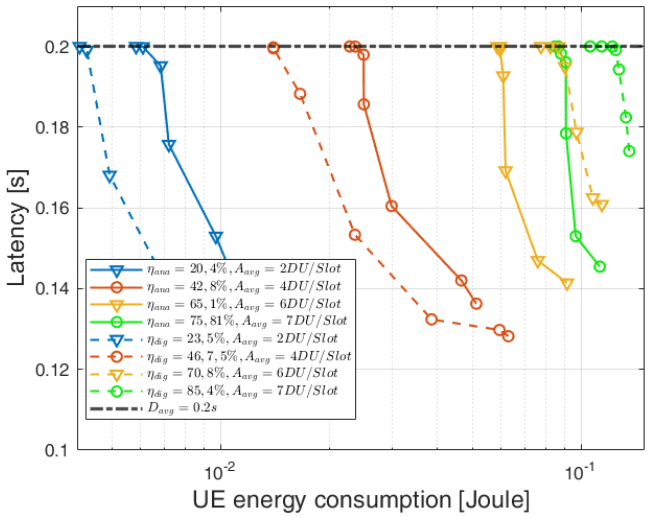

Fig. 3. Energy/Latency trade-off for different loading factors ($n_q = 8bit$).

which, depending on the image arrival rates $A(t)$, measure how much the systems are loaded (in image/seconds), with respect to the maximum load they would be capable to handle using the maximum bandwidth B_{max} and power P_{max}, with $C_{max}(t)$ denoting the Shannon channel capacity. Fig. 3 shows the accuracy/latency trade-offs for different arrivals rates A_{avg}, reporting also the *average* traffic load experienced at convergence by the two resource allocation policies. The curves are explored by increasing the value of the parameter V, from the bottom right to the top-left, which is the optimal point that strictly satisfies the delay constraint, where we computed the two average load factors. Actually, the (ideal) digital system has lower energy expenditure for very low loading factors, i.e., when the system is over-designed resource-wise. Conversely, when B_{max} and P_{max} are better fitted to the actual traffic needs, the analog solution outperforms the digital one. Intuition suggests that when bandwidth and power resources are abundant, *ideal* digital-PAM with zero-BER can outperform analog-PAM, because it is capable to clean the AWGN noise at the receiver side, without suffering the accuracy degradation due the (small) quantization noise introduced at the transmitter. This advantage however, may significantly reduce in practical systems, with non-negligible BER that, for stringent-latency constraints, may request sub-optimal short channel codes. This aspect deserves to be more deeply investigated for MQAM-PAM equipped with PCA-based compression, which should suffer less than JPEG for residual BER, when transmission and computations resources are managed by proper optimization policies, similarly to what has been done in [13].

V. Conclusion and Future Work

This paper shades some light on analog vs digital GOC systems, taking into account an holistic resource management strategy, for image classification. The results confirm that analog GOC systems are indubitably attractive for their easier implementation and capability to work at very low SNRs, and make better use of resources if the system is pushed towards its capacity limit. Future works will consider BER-aware resource management [13], optimal selection of the number of PCA features (and quantization bits), as well as frequency selective fading channels, and multi-user/multi-server management.

References

[1] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez *et al.*, "6g: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 42–50, 2019.

[2] W. Jiang *et al.*, "The road towards 6g: A comprehensive survey," *IEEE Open Jour. of the Comm. Soc.*, vol. 2, pp. 334–366, 2021.

[3] C. Chaccour, W. Saad, M. Debbah *et al.*, "Less data, more knowledge: Building next generation semantic communication networks," *arXiv preprint arXiv:2211.14343*, 2022.

[4] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *arXiv preprint arXiv:2102.04170*, 2021.

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

[6] ——, "Wireless edge machine learning: Resource allocation and trade-offs," *IEEE Access*, vol. 9, pp. 45 377–45 398, 2021.

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

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

[9] G. Chechik *et al.*, "Information bottleneck for gaussian variables," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[10] F. Pezone *et al.*, "Goal-oriented communication for edge learning based on the information bottleneck," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2022, pp. 8832–8836.

[11] F. Binucci, P. Banelli, P. Di Lorenzo *et al.*, "Adaptive resource optimization for edge inference with goal-oriented communications," *EURASIP Jour. on Advan. in Sig. Proc.*, vol. 2022, no. 1, p. 123, 2022.

[12] ——, "Dynamic resource allocation for multi-user goal-oriented communications at the wireless edge," in *2022 30th European Sig. Proc. Conf. (EUSIPCO)*. IEEE, 2022, pp. 697–701.

[13] F. Binucci and P. Banelli, "Ber-aware dynamic resource management for edge-assisted goal-oriented communications," in *2023 Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE Jour. on Selec. Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2020.

[15] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. on Cognitive Communic. and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[16] S. Xie, Y. Wu, S. Ma *et al.*, "Robust information bottleneck for task-oriented communication with digital modulation," 2022.

[17] T.-Y. Tung, D. B. Kurka, M. Jankowski *et al.*, "Deepjscc-q: Constellation constrained deep joint source-channel coding," *IEEE Jour. on Selec. Areas in Inf. Theory*, 2023.

[18] D. Huang, F. Gao, X. Tao *et al.*, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE Jour. on Selected Areas in Communications*, vol. 41, no. 1, pp. 55–71, 2023.

[19] A. Goldsmith, *Wireless communications*. Cambridge univ.press, 2005.

[20] H. Zou, C. Zhang, S. Lasaulce *et al.*, "Goal-oriented quantization: Analysis, design, and application to resource allocation," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 42–54, 2023.

[21] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 13, no. 2-3, pp. 203–221, 1996.

[22] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[23] A. F. Molisch, "Statistical description of the wireless channel," 2011.

[24] S. Sun, T. S. Rappaport, S. Rangan *et al.*, "Propagation path loss models for 5g urban micro-and macro-cellular scenarios," in *83rd IEEE Vehic. Tech. Conf.(VTC Spring)*, 2016, pp. 1–6.

[25] J. Stallkamp *et al.*, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.