Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه

Statistical Federated Learning for Beyond 5G SLA-Constrained RAN Slicing

Hatim Chergui, *Member, IEEE*, Luis Blanco, *Member, IEEE*, and Christos Verikoukis, *Senior Member, IEEE*

***Abstract*— A key enabler for both scalability and sustainability in beyond 5G (B5G) network slicing consists on minimizing the exchange of raw monitoring data across different domains. This is achieved by bringing the analysis functions closer to the data collection points. To this end, we introduce in this paper *statistical federated learning* (SFL) provisioning models that can learn over a live network non independent identically distributed (non-IID) datasets in an offline fashion while respecting slice-level service level agreement (SLA) long-term statistical constraints. Specifically, we consider three resource SLA metrics, namely, *cumulative distribution function* (CDF), *Q-th percentile* and *maximum/minimum bounds*. These metrics are dataset-dependent and non-convex non-differentiable and, to sidestep the inaccuracy of settling only for surrogates, we propose a novel formulation that jointly considers the statistical objective and constraints as well as their smooth approximation using the *proxy-Lagrangian* framework, which we solve via a non-zero sum two-player game strategy. Numerical results on various slice-level resources show that SFL enables SLA enforcement while significantly reducing the overhead compared to both state-of-the-art FedAvg and centralized constrained deep learning schemes. Finally, we provide an analysis for the lower bound of the so-called *reliable convergence probability* in the SFL setup.**

***Index Terms*— B5G, cloud-RAN, dynamic slicing, game theory, proxy-Lagrangian, resource allocation, SLA, statistical federated learning, violation rate.**

I. Introduction

BEYOND 5G systems will leverage the recent advances in network virtualization, softwarization and programmability to enable network slicing technology. Unlike existing architectures, network slicing paves the way for vertical tenants to rent and manage tailored and isolated logical networks—or slices—on top of the same physical infrastructure [1]. The full isolation of slices, however, may induce a high cost in terms of efficiency, and urges the adoption of dynamic orchestration of resources, at least at network edge, to benefit from resource multiplexing [2], [3]. On the other hand, a notion of SLA is also required to guarantee the target quality of service (QoS) levels while conveying network slices on top of a physical network. In this respect, each SLA includes specific metrics that are used to assess slice performance, which means that the management of SLA should be automated for the sake of accountability of various network conditions and variety of user patterns over different slices [4]. In this regard, artificial intelligence (AI) techniques are expected to be the cornerstone in the automation of dynamic network slicing resource allocation as well as the related SLA. Specifically, in ETSI's standardized zero-touch architecture [6], each network domain is endowed with a data collection element that feeds local domain analytics and intelligence entities, while the central end-to-end domain may play the role of a coordinator.

To achieve both the scalability and sustainability of network slicing, AI analysis algorithms should be brought closer to the distributed monitoring points across the network in such a way the raw data exchange is dramatically reduced and only some AI models parameters are transported for coordination or collaboration intents. In such a case, federated learning (FL)-based analysis is an interesting solution to solve this two-fold problem since it enables i) learning locally at each network element, either virtual or physical, while exchanging only the weights of its model with the aggregation server and, ii) resolving the lack or poor distribution of the local datasets by leveraging the knowledge of the other elements taking part in the FL task [5]. The open challenge in this case is to i) guarantee that the outputs of the offline learned models are respecting some preset statistical distribution (such as SLA) when deployed in a test scenario and, ii) ensure the convergence of these models under live network non-IID datasets.

A. Related Work

In [8], a distributed framework called federated-orchestrator (F-orchestrator) has been proposed, which can coordinate slices' spectrum and computational resources via a distributed resource allocation algorithm termed *Alternating Direction Method of Multipliers with Partial Variable Splitting* (DistADMM-PVS). In particular, it has been proved that DistADMM-PVS minimizes the average service response time of the entire network with guaranteed worst-case performance for the different services, when the coordination between the F-orchestrator and base stations (BSs) is perfectly synchronized.

In [9] and [10], the authors have analyzed in closed-form the convergence rates of federated learning (FL) algorithms, based on which they have derived a link between network resources (e.g., delay, computation and transmission energies) and FL

Manuscript received October 9, 2020; revised May 31, 2021; accepted August 20, 2021. Date of publication September 9, 2021; date of current version March 10, 2022. This work was supported in part by the Research Project MonB5G under Grant 871780, in part by the Research Project Suport als Grups de Recerca (SGR) under Grant 2017 SGR 1195, and in part by the Research Project ZEROTO6G. The associate editor coordinating the review of this article and approving it for publication was A. El Gamal. *(Corresponding author: Hatim Chergui.)*

The authors are with the Catalan Telecommunications Technology Center (CTTC), 08860 Barcelona, Spain (e-mail: hatim.chergui@cttc.es; luis.blanco@cttc.es; cveri@cttc.es).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3109377.

Digital Object Identifier 10.1109/TWC.2021.3109377

1536-1276 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه

performance and tried to jointly optimize the underlying learning time under network resource constraints, capturing thereby the trade-off of latency and energy consumption for FL.

In [11], the authors have presented two centralized scheduling algorithms to allocate resources in network slicing systems. The algorithms take into account the latency requirement for different services and the SLA requirement in terms of minimal demand that has to be allocated to each tenant. While the first algorithm considers only the tenant priority, the second one takes into account also the availability rate aiming at yielding time-fair allocations.

In [12], we have proposed centralized deep neural networks (DNNs)-based models to estimate the required resources per slice, while not violating two service level agreement (SLA), namely, violation rate-based SLA and resource bounds-based SLA. This is achieved by integrating dataset-dependent generalized non-convex constraints into the DNN offline optimization.

In [13], DeepCog has been introduced as a centralized data analytics tool for the cognitive management of resources in sliced 5G systems. DeepCog forecasts the capacity required to accommodate future traffic demands within individual network slices while accounting for the operator's desired balance between resource overprovisioning and SLA violations. To this end, DeepCog leverages a convolutional neural network (CNN)-based model trained with a custom loss function.

In [14], GREET has been proposed as a slice-based resource allocation framework which relies on a centralized constrained resource allocation game where slices can unilaterally optimize their allocations under some constraints. This guarantees that slices are assigned the share of network resources as specified in their SLAs.

A centralized slice-aware radio resource management (RRM) scheme has been presented in [15], wherein a network entity called mapping layer keeps track of the load and performance KPIs of different slices and, according to their SLAs, a centralized algorithm tries to minimize the deviations from target KPIs by fine-tuning weighting parameters of the packet scheduler so that the SLA targets for slices are fulfilled.

While we notice that many efforts have been deployed to consider SLA in network slicing resource allocation schemes, we remark that there is no decentralized approach directly integrating practical long-term SLA constraints into the optimization of slices' resource provisioning models.

B. Contributions

In this paper, our contribution is multi-fold:

- By invoking live network key performance indicators (KPIs) non-IID datasets, we introduce statistical federated learning (SFL) resource allocation models that can learn over a data distribution in an offline fashion while fulfilling some predefined local long-term SLA constraints. In particular, we consider three SLA statistical metrics defined over a time window, namely, *resource cumulative distribution function*, *Q-th resource percentile* and *maximum/minimum resource bounds*.
- Instead of optimizing with respect to only some surrogates of the non-convex non-differentiable original statistical metrics—which might not accurately verify the original constraints—we jointly optimize the loss function, the original as well as their smoothed proxy surrogates using the *proxy-Lagrangian* framework via a non-zero sum two-player game strategy that is integrated with the FL setup
- By evaluating our schemes over separate test dataset, we show that the proposed decentralized resource allocation approach generalizes well and enables SLA enforcement compared to unconstrained FedAvg [16] while also significantly reducing the communication overhead compared to a centralized constrained deep neural network scheme [12] for equal loss.
- Using reliability theory, we provide a closed-form analysis for the lower bound of the so-called *reliable convergence probability* where both, the SLA fulfillment and the convergence rate of the federated models, are jointly characterized.

TABLE I
NOTATIONS

Notation	Description
$\odot$	Element-wise multiplication
$S_\mu(\cdot)$	Logistic function with steepness μ
$f_Q(\cdot)$	Q-th percentile function
$F(\cdot)$	CDF
$\tilde{F}(\cdot)$	Complementary CDF
L	Number of local epochs
T	Number of FL rounds
$\mathcal{D}_{k,n}$	Dataset at MS (k,n)
$D_{k,n}$	Dataset size
$\ell(\cdot)$	Loss function
$\mathbf{W}_{k,n}^{(t)}$	Local weights of AE (k,n) at round t
$\mathbf{x}_{k,n}$	Input features
$\hat{y}_{k,n}^{(i)}$	Allocated resource by AE (k,n)
α_n	Resource lower-bound for slice n
β_n	Resource upper-bound for slice n
γ_n	Threshold of CDF-based SLA for slice n
π_n	Threshold of Q-th percentile-based SLA for slice n
$\lambda_{(\cdot)}$	Lagrange multipliers
R_λ	Lagrange multiplier radius
$\mathcal{L}_{(\cdot)}$	Lagrangian with respect to $(\cdot)$

C. Notations

We summarize the notations used throughout the paper in Table I.

II. Network Architecture and Datasets

As depicted in Fig. 1, we consider a beyond 5G RAN architecture under the central unit (CU)/distributed unit (DU) functional split, wherein each transmission/reception point (TRP) is co-located with its DU which is connected to the corresponding CU by a fronthaul link. In this respect, each CU $k\,(k=1,\ldots,K)$ runs as a virtual network function (VNF) on top of a commodity hardware, and performs slice-level RAN key performance indicators data collection via a *monitoring system* (MS) as well as implements AI-enabled slice resource allocation through the so-called *analytics engine* (AE). For each CU k and slice $n\,(n=1,\ldots,N)$, MS (k,n) has a local

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه

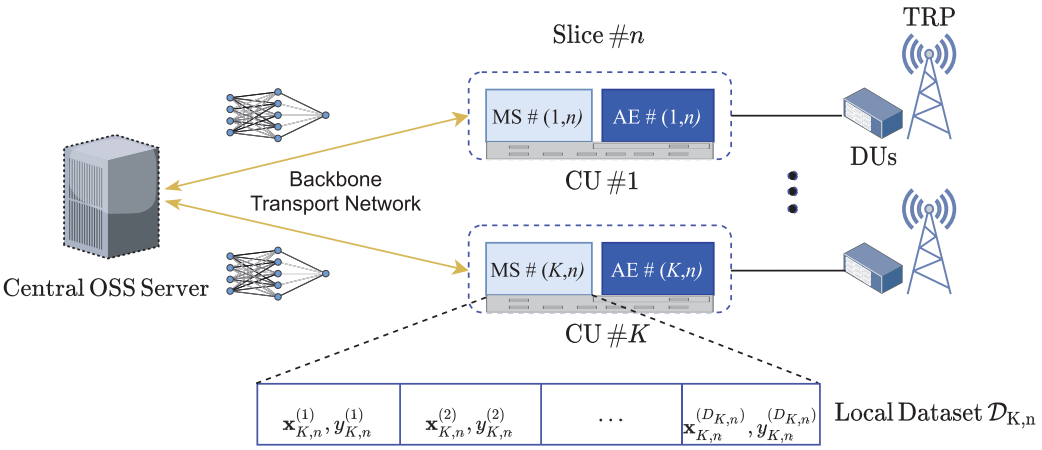


Fig. 1. Network architecture with decentralized MS/AE.

TABLE II
DATASET FEATURES AND OUTPUT

Feature	Description
OTT Traffics	Apple, Facebook, Facebook Messages, Facebook Video, Instagram, NetFlix, HTTPS, QUIC, Whatsapp, and Youtube
CQI	Channel quality indicator
MIMO Full-Rank	MIMO full-rank usage (%)
# Users	Downlink Average active users

Output	Description
DLPRB	Traffic volume
CPU Load	CPU resource consumption (%)
RRC Connected Users	Number of RRC users licenses

dataset $\mathcal{D}_{k,n}$ of size $D_{k,n}$ that is generally small and non-exhaustive. Therefore, the corresponding local AE participates in a federated learning task—to accurately train its resource provisioning model—and is thereby connected to a central operational subsystem (OSS) server that plays the role of model aggregator without having access to the raw datasets.

As summarized in Table II, by leveraging both Huawei PRS OSS tool and Tektronix probes, we have collected the datasets from a live LTE-advanced network of 3200 TRPs with a granularity of 1 hour, and have a size $D_{k,n} = 1000$ each. The considered TRPs cover areas with different traffic behaviors—both in space and time—which tightly depend on the heterogeneous users distribution and profiles in each context (e.g., residential zones, business zones, entertainment events, ...). On the other hand, the radio KPIs are correlated with the time-varying channel conditions. These realistic datasets are therefore non-IID per se. Each local dataset includes, as input features, the hourly traffics of the main over-the-top (OTT) applications, channel quality indicator (CQI), and multiple-input multiple-output (MIMO) full-rank usage. The supervised output KPI might be either the number of occupied downlink (DL) physical resource blocks (PRBs), or the central processing unit (CPU) load or the number of RRC connected users. Although the PRBs, CPU usage and RRC users are also measured, but they result from the eNodeB built-in scheduling algorithms, such as Proportional Fairness (PF), Round Robin (RR) or maximum CQI, which are different from our AI-based approach. Therefore, the role of our novel FL strategies is to predict these resources to fulfill some preset statistical SLA constraints.

Once the slices are defined, the traffic of the underlying OTTs is summed to yield the traffic per slice. In this regard, for each dataset $\mathcal{D}_{k,n} = \{\mathbf{x}_{k,n}^{(i)}, y_{k,n}^{(i)}\}_{i=1}^{D_{k,n}}$, $\mathbf{x}_{k,n}^{(i)}$ stands for the input features vector while $y_{k,n}^{(i)}$ represents the corresponding output.

III. Slice-Level Federated Resource Allocation Under Statistical SLA Constraints

In this section, we describe in detail the analytics engine federated learning-based resource allocation under practical SLA constraints. Specifically, for each local AE (k, n), the predicted amount of resources $\hat{y}_{k,n}^{(i)}$, $(i = 1, \ldots, D_{k,n})$, should minimize the main loss function with respect to the ground truth $y_{k,n}^{(i)}$, while also respecting some long-term statistical constraints defined over its $D_{k,n}$ samples. The optimized local weights at round t, $\mathbf{W}_{k,n}^{(t)}$, are sent to the OSS server that generates a global FL model for slice n as,

$$\mathbf{W}_n^{(t+1)} = \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} \mathbf{W}_{k,n}^{(t)}, \tag{1}$$

where $D_n = \sum_{k=1}^{K} D_{k,n}$ is the total data samples of all datasets related to slice n. The server then broadcasts the global model (1) to all K AEs that use it to start the next round of local optimization. The proposed statistical FL leverages a two-player game strategy to jointly optimize over the objective and original constraints as well as their smoothed surrogates as summarized in Fig. 2 and detailed in the sequel for various SLA metrics.

A. Cumulative Distribution Function SLA

According to the SLA established between slice n tenant and the physical operator, any assigned resource to the tenant should not exceed a range $[\alpha_n, \beta_n]$ with a probability higher than an agreed threshold γ_n. This translates into learning the resource allocation model under empirical cumulative density

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه
puyamoshaver@gmail.com | شماره تماس: ۰۹۳۵۴۹۴۱۴۳۴

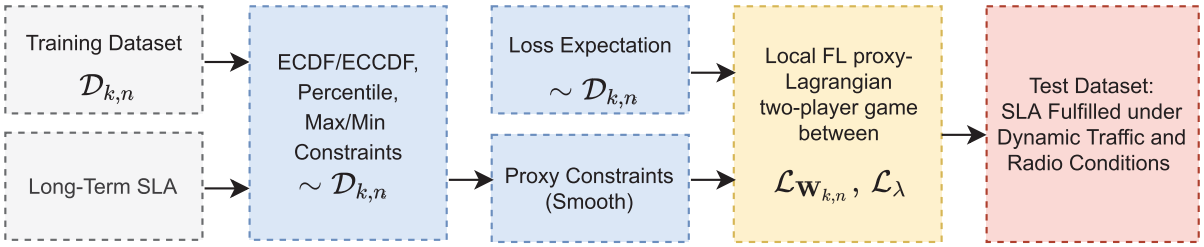


Fig. 2. Analytic engine local offline datatset-dependent statistical SLA learning.

function constraints, which amounts to solving the following local optimization task at FL round $t\,(t=0,\ldots,T-1)$,

$$\min_{\mathbf{W}_{k,n}^{(t)}} \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right), \tag{2a}$$

$$\text{s.t. } F_{\mathbf{x}_{k,n}\sim\mathcal{D}_{k,n}}(\alpha_n) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \mathbb{1}\left(\hat{y}_{k,n}^{(i)} < \alpha_n\right) \leq \gamma_n, \tag{2b}$$

$$\tilde{F}_{\mathbf{x}_{k,n}\sim\mathcal{D}_{k,n}}(\beta_n) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \mathbb{1}\left(\hat{y}_{k,n}^{(i)} > \beta_n\right) \leq \gamma_n, \tag{2c}$$

where $\ell(\cdot)$ is the squared error loss function, $\mathbb{1}(\cdot)$ stands for the indicator function, while $F_{\mathbf{x}_{k,n}\sim\mathcal{D}_{k,n}}$ and $\tilde{F}_{\mathbf{x}_{k,n}\sim\mathcal{D}_{k,n}}$ are the empirical CDF and complementary CDF (CCDF) over dataset $\mathcal{D}_{k,n}$, respectively. They are both linear combinations of indicator functions and, therefore, not even subdifferentiable w.r.t. $\mathbf{W}_{k,n}^{(t)}$. Resolving this issue by replacing the constraints with differentiable surrogates introduces a new difficulty: solutions to the resulting problem will satisfy the surrogate constraints, rather than the original ones. To sidestep this hindrance, let us consider the functions Φ_1 and Φ_2 defined as,

$$\Phi_1(\mathbf{W}_{k,n}^{(t)}) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \mathbb{1}\left(\hat{y}_{k,n}^{(i)} < \alpha_n\right) - \gamma_n, \tag{3}$$

$$\Phi_2(\mathbf{W}_{k,n}^{(t)}) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \mathbb{1}\left(\hat{y}_{k,n}^{(i)} > \beta_n\right) - \gamma_n, \tag{4}$$

and let Ψ_1 and Ψ_2 be sufficiently-smooth surrogates of Φ_1 and Φ_2 [17] verifying

$$\Psi_1(\mathbf{W}_{k,n}^{(t)}) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} S_\mu\left(\alpha_n - \hat{y}_{k,n}^{(i)}\right) - \gamma_n \leq 0, \tag{5}$$

$$\Psi_2(\mathbf{W}_{k,n}^{(t)}) = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} S_\mu\left(\hat{y}_{k,n}^{(i)} - \beta_n\right) - \gamma_n \leq 0, \tag{6}$$

where L_μ stands for the Logistic function with steepness parameter μ, i.e.,

$$S_\mu(\theta) = \frac{1}{1+e^{-\mu\theta}}. \tag{7}$$

Instead of optimizing with respect convexified surrogates only (using e.g., the convex-concave procedure (CCP) [18]), which might lead to inaccurate results, we formulate the local problem (2) by rather invoking the so-called *proxy Lagrangian* framework [19], where we jointly optimize over two Lagrangians, $\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}$ that includes the original loss and the smoothed surrogates, while $\mathcal{L}_\lambda$ consists of the original constraints, i.e.,

$$\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right) + \lambda_1\Psi_1\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_2\Psi_2\left(\mathbf{W}_{k,n}^{(t)}\right), \tag{8a}$$

$$\mathcal{L}_\lambda = \lambda_1\Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_2\Phi_2\left(\mathbf{W}_{k,n}^{(t)}\right), \tag{8b}$$

where the local optimization task is written as,

$$\min_{\mathbf{W}_{k,n}\in\Delta} \max_{\lambda,\,\|\lambda\|\leq R_\lambda} \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} \tag{9a}$$

$$\max_{\lambda,\,\|\lambda\|\leq R_\lambda} \min_{\mathbf{W}_{k,n}\in\Delta} \mathcal{L}_\lambda, \tag{9b}$$

whose optimization turns out to be a non-zero-sum two-player game in which the $\mathbf{W}_{k,n}^{(t)}$-player aims at minimizing $\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}$, while the λ-player wishes to maximize $\mathcal{L}_\lambda$ [20, Lemma 8]. While optimizing the first Lagrangian w.r.t. $\mathbf{W}_{k,n}$ requires differentiating the constraint functions $\Psi_1(\mathbf{W}_{k,n}^{(t)})$ and $\Psi_2(\mathbf{W}_{k,n}^{(t)})$, to differentiate the second Lagrangian w.r.t. λ we only need to evaluate $\Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right)$ and $\Phi_2\left(\mathbf{W}_{k,n}^{(t)}\right)$. Hence, a surrogate is only necessary for the $\mathbf{W}_{k,n}$-player; the λ-player can continue using the original constraint functions. Via Lagrange multipliers, the λ-player chooses how much to weigh the proxy constraint functions, but does so in such a way as to satisfy the original constraints, and ends up reaching a nearly-optimal nearly-feasible solution [21]. This approximate equilibrium corresponds to a ν-approximate saddle point of the Lagrangian, which is a pair $\left(\hat{\mathbf{W}}_{k,n}^{(t)}, \hat{\lambda}\right)$, where according to the preset accuracy ν,

$$\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}(\hat{\mathbf{W}}_{k,n}^{(t)}, \hat{\lambda}) \leq \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}(\mathbf{W}_{k,n}^{(t)}, \hat{\lambda}) + \nu,\ \mathbf{W}_{k,n}^{(t)} \in \Delta \tag{10a}$$

$$\mathcal{L}_\lambda(\hat{\mathbf{W}}_{k,n}^{(t)}, \hat{\lambda}) \geq \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}(\mathbf{W}_{k,n}^{(t)}, \lambda) - \nu,\ \|\lambda\| \leq R_\lambda \tag{10b}$$

This federated learning task with the two-player game solution for local problem (2) can be summarized as in Algorithm 1.

Using Google's `tensorflow constrained optimization package` [22], we implement algorithm 1 that uses two different approaches to optimize the Lagrangians at each local epoch $l\ (l = 0,\ldots,L-1)$: a Bayesian optimization oracle for $\mathcal{L}_{\mathbf{W}_{k,n,l}}$ and projected gradient ascent with learning rate η_λ for $\mathcal{L}_\lambda$. A definition of the oracle $\mathcal{O}_\delta$ is given as follows:

Definition 1 (Approximate Bayesian Optimization Oracle): A δ-approximate Bayesian optimization oracle is a routine $\mathcal{O}_\delta$

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه

Algorithm 1 Federated learning with local proxy-Lagrangian two-player game for slice n.

Input: R_λ, η_λ, T, L. `# See Table I`
OSS server initializes $\mathbf{W}_n^{(0)}$ with random Gaussian weights and broadcasts it to local AEs.
for $t = 0, \ldots, T-1$ **do**
  parallel for $k = 1, \ldots, K$
  Initialize $M =$ `num_constraints` and $\mathbf{W}_{k,n,0} = \mathbf{W}_n^{(t)}$
  Initialize $\mathbf{A}^{(0)} \in \mathbb{R}^{(M+1)\times(M+1)}$ with $\mathbf{A}^{(0)}_{m',m} = 1/(M+1)$
  for $l = 0, \ldots, L-1$ **do**
    Let $\lambda^{(l)}$ be the top eigenvector of $A^{(l)}$
    `# Oracle optimization`
    Let $\hat{\mathbf{W}}_{k,n,l} = \mathcal{O}_\delta\left(\mathcal{L}_{\mathbf{W}_{k,n,l}}(\cdot, \hat{\lambda}^{(l)})\right)$
    Let $\Delta_\lambda^{(l)}$ be a gradient of $\mathcal{L}_\lambda(\hat{\mathbf{W}}_{k,n,l}, \lambda^{(l)})$ w.r.t. λ
    `# Exponentiated gradient ascent`
    Update $\tilde{\mathbf{A}}^{(l+1)} = \mathbf{A}^{(l)} \odot \cdot \exp \eta_\lambda \Delta_\lambda^{(l)}(\lambda^{(l)})$
    `# Colunm-wise normalization`
    $\mathbf{A}_m^{(l+1)} = \tilde{\mathbf{A}}_m^{(l+1)} / \left\|\mathbf{A}_m^{(l+1)}\right\|_1, \; m = 1, \ldots, M+1$
    $\nu_l = \left|\mathcal{L}_\lambda(\hat{\mathbf{W}}_{k,n}^{(t)}, \hat{\lambda}) - \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}(\mathbf{W}_{k,n}^{(t)}, \lambda)\right|$
    if $\nu_l \le \nu \cup l = L-1$ **then**
      $L^\star = l$;
      Break;
    end
  end
  return $\hat{\mathbf{W}}_{k,n}^{(t)} = \frac{1}{L^\star}\sum_{l=0}^{L-1} \hat{\mathbf{W}}_{k,n,l}$
  Each local AE (k,n) sends $\hat{\mathbf{W}}_{k,n}^{(t)}$ to the OSS server.
  end parallel for
  return $\mathbf{W}_n^{(t+1)} = \sum_{k=1}^K \frac{D_{k,n}}{D_n} \hat{\mathbf{W}}_{k,n}^{(t)}$
  and broadcasts the value to all local AEs.
end

that given a loss function/Lagrangian $\mathcal{L}$, returns the quasi-optimal weights $\mathbf{W}_{k,n,l}$ such that

$$\mathcal{L}\left(\mathcal{O}_\delta\left(\mathcal{L}\right)\right) \le \inf_{\mathbf{W}^\star_{k,n,l}} \mathcal{L}\left(\mathbf{W}^\star_{k,n,l}\right) + \delta. \tag{11}$$

Note that $\lambda_1, \lambda_2 \le R_\lambda$, where R_λ represents the radius of Lagrange multipliers; introduced as a hyperparameter that can be fine-tuned to control the level of dependency to the constraints.

B. Q-Th Percentile SLA

Practical SLAs can also be defined in terms of percentiles. In this respect, the physical operator might agree with the tenant of slice n that e.g., the Q-th percentile of a specific resource shall be lower than a threshold π_n to ensure isolation. At each AE (k,n), this consists on allocating resources according to the local optimization task,

$$\min_{\mathbf{W}_{k,n}^{(t)}} \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right), \tag{12a}$$

$$\text{s.t. } f_Q\left(\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right) \le \pi_n, \tag{12b}$$

where $f_Q(\cdot)$ is the Q-th percentile statistic over set $\mathcal{Y}_{k,n} = \left\{\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right\}$. It can be expressed in an explicit way by first sorting the elements of $\mathcal{Y}_{k,n}$ in the ascending order, i.e.,

$$\mathcal{Y}_{k,n} = \{z_1, \ldots, z_{D_{k,n}} \mid z_i < z_{i+1}\}, \tag{13}$$

and then calculating it as follows,

$$f_Q\left(\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right) = z_j, \; j = \lfloor * \rfloor \frac{Q \times (D_{k,n}+1)}{100}. \tag{14}$$

To solve problem (12), let us consider function Φ_3,

$$\Phi_3\left(\mathbf{W}_{k,n}^{(t)}\right) = f_Q - \pi_n. \tag{15}$$

A smooth approximation Ψ_3 of (15) can be obtained by invoking the so-called *smoothed empirical percentile* [23], i.e.,

$$\Psi_3\left(\mathbf{W}_{k,n}^{(t)}\right) = \tilde{f}_Q - \pi_n \le 0, \tag{16}$$

where,

$$\tilde{f}_Q = (1-h)\, z_j + h z_{j+1}, \tag{17}$$

and $h = \frac{Q(D_{k,n}+1)}{100} - j$. At this level, we form the two-player Lagrangians similarly to the previous subsection, i.e.,

$$\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right) + \lambda_3 \Psi_3\left(\mathbf{W}_{k,n}^{(t)}\right) \tag{18a}$$

$$\mathcal{L}_\lambda = \lambda_3 \Phi_3\left(\mathbf{W}_{k,n}^{(t)}\right), \tag{18b}$$

and run Algorithm 1 to optimize them.

C. Maximum/Minimum SLA

Another type of statistical SLA consists on thresholds imposed to the maximum and minimum resources granted to each slice over observations $\left\{\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right\}$. Similarly to problem (2), we write this federated learning local optimization task at AE (k,n) as follows:

$$\min_{\mathbf{W}_{k,n}^{(t)}} \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right), \tag{19a}$$

$$\text{s.t. } \Phi_4\left(\mathbf{W}_{k,n}^{(t)}\right) = \alpha_n - \min_i \hat{y}_{k,n}^{(i)} \le 0, \tag{19b}$$

$$\Phi_5\left(\mathbf{W}_{k,n}^{(t)}\right) = \max_i \hat{y}_{k,n}^{(i)} - \beta_n \le 0. \tag{19c}$$

To derive the proxy constraints, we seek smooth upper bounds on functions Φ_4 and Φ_5. In this respect, we invoke the smooth maximum and minimum functions [24] that read,

$$\sigma_{\max}\left(\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right) = \log\left(\sum_{i=1}^{D_{k,n}} \exp \hat{y}_{k,n}^{(i)}\right), \tag{20}$$

$$\sigma_{\min}\left(\hat{y}_{k,n}^{(1)}, \ldots, \hat{y}_{k,n}^{(D_{k,n})}\right) = -\log\left(\sum_{i=1}^{D_{k,n}} \exp -\hat{y}_{k,n}^{(i)}\right). \tag{21}$$

We then express the proxy constraints as,

$$\Psi_4\left(\mathbf{W}_{k,n}^{(t)}\right) = \alpha_n - \sigma_{\min} \le 0, \tag{22a}$$

$$\Psi_5\left(\mathbf{W}_{k,n}^{(t)}\right) = \sigma_{\max} - \beta_n \le 0. \tag{22b}$$

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه
puyamoshaver@gmail.com | شماره تماس : ۰۹۳۵۴۹۴۱۴۳۴

Finally, we write the two Lagrangians,

$$\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(\mathbf{y}_{k,n}^{(i)}, \hat{\mathbf{y}}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right) + \lambda_4 \Psi_4\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_5 \Psi_5\left(\mathbf{W}_{k,n}^{(t)}\right), \quad (23a)$$

$$\mathcal{L}_{\lambda} = \lambda_4 \Phi_4\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_5 \Phi_5\left(\mathbf{W}_{k,n}^{(t)}\right), \quad (23b)$$

and use Algorithm 1 to optimize them similarly to the previous subsection.

IV. Federated Learning Convergence Analysis

In this section, we analyze the convergence probability of the SLA-constrained federated learning models. In this intent, a closed-form expression for the lower bound of the convergence probability is derived, reflecting the effects of the Lagrange multiplier radius, the optimization oracle error and the violation rate.

Theorem 1 (Convergence Analysis of the SLA-Constrained Federated Learning): Consider that the federated learning fails to fulfill the constraints with average violation rate $0 < \nu < 1$, and follows a geometric failure model. It is also assumed that $\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}$ is optimized using an oracle $\mathcal{O}_\delta$ with error δ, and let R_λ and $B_{k,n}$ stand for the Lagrange multipliers radius and the upper bound on the norm of subgradient $\nabla\mathcal{L}_\lambda\left(\mathbf{W}_{k,n}^{(t)}, \lambda^{(t)}\right)$, respectively. Then, the federated learning convergence probability satisfies,

$$\mathbf{Pr}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^\star\right) -, \inf_{\mathbf{W}_n^\star} \mathcal{L}_\lambda\left(\mathbf{W}_n^\star, \lambda^{(t)}\right)\right) < \epsilon\right] \geq \Delta\left(\nu, \epsilon\right), \quad (24)$$

where

$$\Delta\left(\nu, \epsilon\right) = 1 - \frac{\nu}{1 + (\nu - 1)\exp\frac{-(D_n\epsilon)^2}{2(2R_\lambda \sum_{k=1}^{K} D_{k,n}B_{k,n} + D_n\delta)^2}}. \quad (25)$$

Proof: First, by means of the subgradient inequality we have at round t,

$$\mathcal{E}^{(t)} = \mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^*\right) - \mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^{(t)}\right) \leq \langle \nabla\mathcal{L}_\lambda^{(t)}, \lambda^* - \lambda^{(t)}\rangle. \quad (26)$$

Using Cauchy-Schwarz inequality, we get,

$$\mathcal{E}^{(t)} \leq \left\|\nabla\mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^{(t)}\right)\right\| \left\|\lambda^\star - \lambda^{(t)}\right\|. \quad (27)$$

By recalling the federated learning aggregation (1), we can write

$$\nabla\mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^{(t)}\right) = \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} \nabla\mathcal{L}_\lambda\left(\mathbf{W}_{k,n}^{(t)}, \lambda^{(t)}\right). \quad (28)$$

Therefore, from (27) and (28) and by invoking the triangle inequality we have,

$$\mathcal{E}^{(t)} \leq \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} \left\|\nabla\mathcal{L}_\lambda\left(\mathbf{W}_{k,n}^{(t)}, \lambda^{(t)}\right)\right\| \left\|\lambda^\star - \lambda^{(t)}\right\| \leq 2R_\lambda \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} B_{k,n}. \quad (29)$$

At this level, let

$$\mathcal{U}^{(t)} = \mathcal{L}_\lambda\left(\mathbf{W}_n^{(t)}, \lambda^\star\right) - \inf_{\mathbf{W}_n^\star} \mathcal{L}_\lambda\left(\mathbf{W}_n^\star, \lambda^{(t)}\right). \quad (30)$$

Combining (29) and (30) with Definition 1 yields,

$$\mathcal{U}^{(t)} \leq 2R_\lambda \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} B_{k,n} + \delta = C. \quad (31)$$

By means of Hoeffding-Azuma's inequality [25], we have

$$\Pr[\frac{1}{T}\sum_{t=1}^{T} \mathcal{U}^{(t)} < \epsilon \mid T = \tau] \geq 1 - \exp -\frac{\tau\varepsilon^2}{2C^2}, , \quad (32)$$

where we consider that the federated learning model is reliable, i.e., respecting the SLA up to and including time $T = \tau$. Therefore, by recalling the geometric failure probability mass function P_τ given by,

$$P_\tau = \nu\left(1 - \nu\right)^\tau, \quad (33)$$

and combining it with (32), yields

$$\Pr[\frac{1}{T}\sum_{t=1}^{T} \mathcal{U}^{(t)} < \epsilon] \geq \sum_{\tau=0}^{+\infty} \nu\left(1 - \nu\right)^\tau \times \left(1 - \exp -\frac{\tau\epsilon^2}{2C^2}\right). \quad (34)$$

Finally, by noticing that $\nu < 1$ and following some algebraic manipulations, we obtain the target result as in (9) and (25). ■

V. Numerical Results

A. Test Dataset, Settings and Baselines

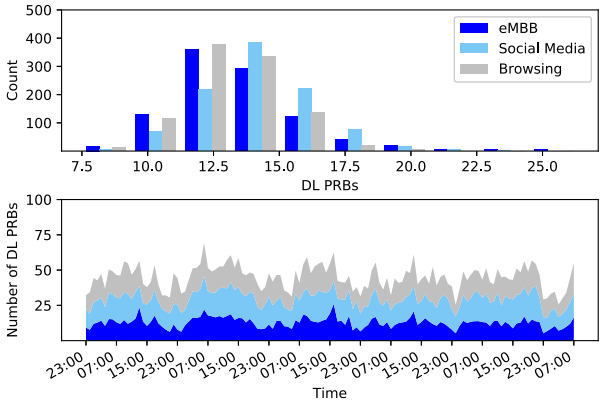
To exemplify the general framework of SFL and shows that it generalizes well after the training, we consider a separate test dataset consisting of the hourly KPIs of three slices, namely:

- **eMBB:** involves NetFlix, Youtube and Facebook Video,
- **Social Media:** includes Facebook, Facebook Messages, Whatsapp and Instagram,
- **Browsing:** encompasses Apple, HTTP and QUIC,

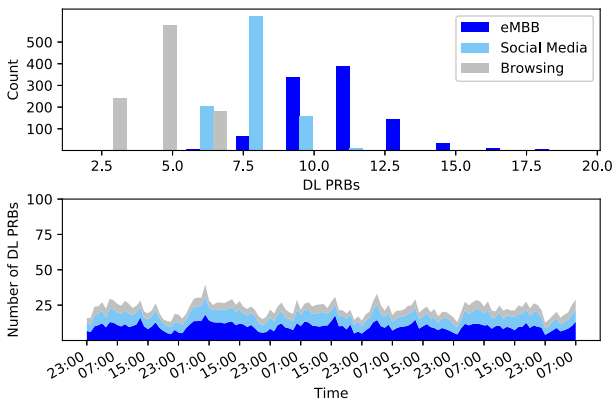
which contains the aggregated traffic per slice, as well as the CQI and MIMO full-rank usage. For the sake of simplicity, we drop index n and use vectors α, β and γ instead. These vectors encompass the resource bounds and thresholds corresponding to the different slices for a particular resource, and can be easily understood from the context. The parameters settings are summarized in Table III. The performance of the proposed solution is assessed using several key performance indicators. The effectiveness of our proposal is compared with a state-of-the-art unconstrained FedAvg FL solution [16] as well as the centralized constrained deep learning (CCL) framework [12].

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

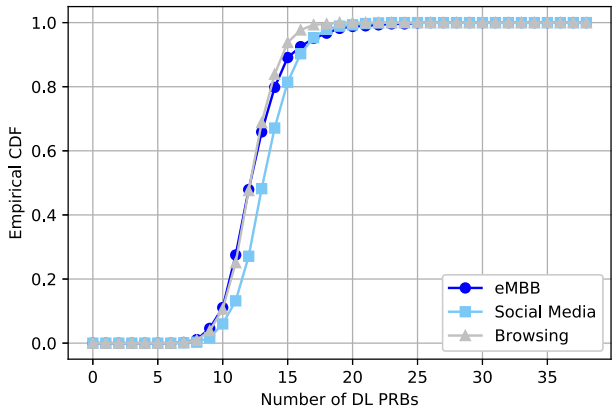
Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه



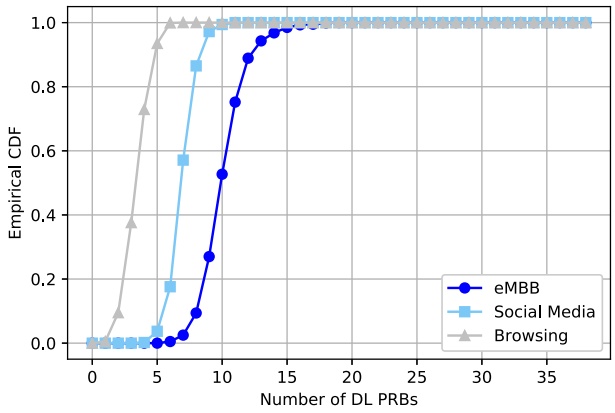
(a) DL PRBs distribution with baseline unconstrained FedAvg



(b) DL PRBs distribution with constrained SFL



(c) DL PRBs CDF with baseline unconstrained FedAvg



(d) DL PRBs CDF with constrained SFL

Fig. 3. DL PRBs distributions using test dataset, with $\alpha = [0, 0, 0]$, $\beta = [15, 10, 10]$ PRBs and $\gamma = [0.01, 0.01, 0.01]$.

TABLE III
SETTINGS

Parameter	Description
N	3 slices
K	200 CUs
L	100 epochs
T	100 rounds
$D_{k,n}$	1000 samples
R_λ	Constrained: 10^{-5}
η_λ	0.02
μ	5

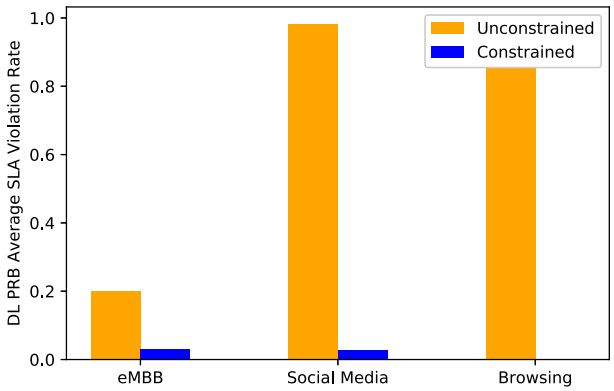


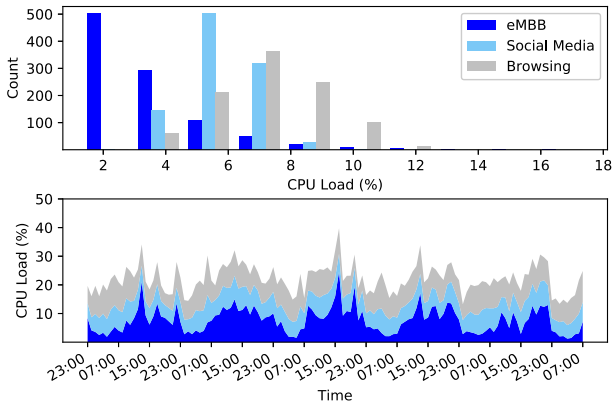
Fig. 4. DL PRBs average violation rate for constrained SFL vs. unconstrained FedAvg using test dataset with $\alpha = [0, 0, 0]$, $\beta = [15, 10, 10]$ PRBs and $\gamma = [0.01, 0.01, 0.01]$.

B. Performance of CDF SLA

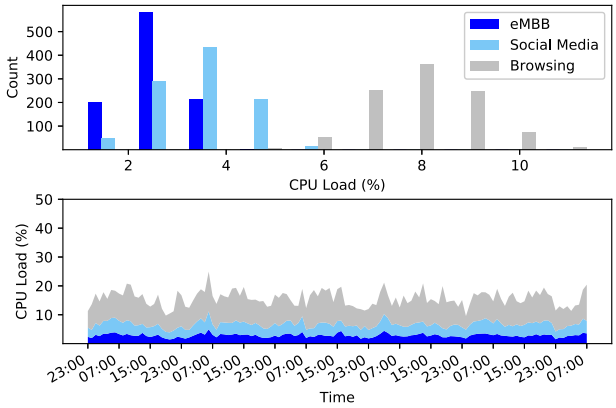
In this scenario, resources at CU-level are dynamically allocated to slices according to their traffic patterns and radio conditions (average CQI, MIMO full-rank usage) as shown in Fig. 3-(a) and 3-(b), while ensuring a long-term isolation via the constraints imposed to the cumulative distribution function (CDF) of the underlying resources. Indeed, in the baseline unconstrained FL, as depicted in Fig. 3-(c), all three slices violate e.g. their upper bounds with a high probability that can be considered as unacceptable by operators and tenants in practice. In contrast, Fig. 3-(d) shows that the number of provisioned DL PRBs is confined within their respective bounds α and β with a probability that reaches 99%. This is achieved by the proposed statistical federated learning (SFL) scheme that trains local resource allocation models to respect preset statistical metrics. As corroborated by Fig. 4, SFL enables the reduction of the DL PRBs average SLA violation rate to the practical value of 1%, and provides an efficient tool to guarantee long-term SLA by overlooking the point-wise resource behavior in favor of a statistical one,

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

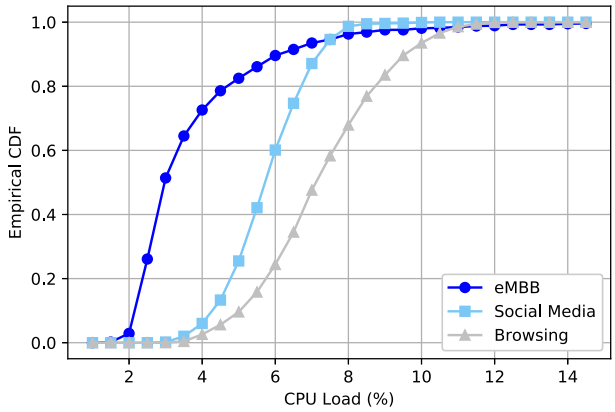
Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com



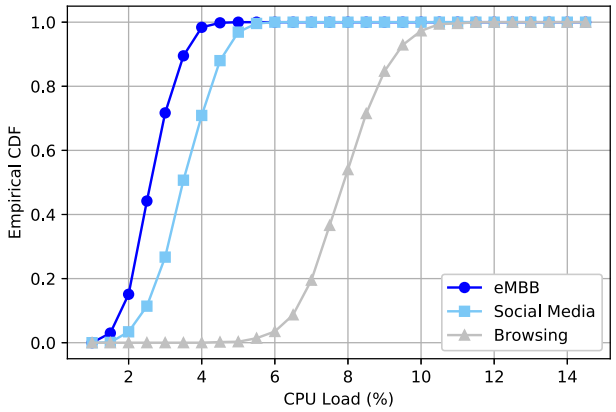
(a) CPU distribution with baseline unconstrained FedAvg



(b) CPU distribution with constrained SFL



(c) CPU CDF with baseline unconstrained FedAvg



(d) CPU CDF with constrained SFL

Fig. 5. CPU load distributions using test dataset, with $\alpha = [0, 0, 0]$, $\beta = [4, 7, 10]$ % and $\gamma = [0.01, 0.01, 0.01]$.

which is adequate for reaching a trade-off between adaptation and isolation in live network operation.

On the other hand, CPU dynamic allocation is required to ensure efficient utilization of the RAN cloud computing capacity and avoid switching-on physical network functions to accommodate VNFs of different slices. The CDF, as a measure of the resource violation rate, can serve to control the long-term CPU load distribution among slices. Indeed, as depicted in Fig. 5-(a), the baseline CPU loads follow the trend of the slices' traffics without respecting the SLA bounds α and β. This behavior is further clarified by the corresponding empirical CDF, in Fig. 5-(c), where it is shown that e.g., eMBB and Social Media slices are breaching the bounds with high probabilities of about 25% and 12%, respectively. This stems from the fact that the baseline FL models cannot learn statistical properties over an observation interval and operate only at sample level. However, in the constrained scenario of Fig. 5-(b), the CPU loads achieve a trade-off between dynamic allocation and long-term statistical SLA. In this case, the eMBB and Social Media CPU loads are confined in the imposed bounds with a high probability of 99% as depicted in Fig. 5-(d). In this respect, Fig. 6 showcases that the CPU load SLA violation rates of the different slices are dramatically reduced in the constrained case and reach the target threshold γ, i.e., 1%, which is an acceptable value for operators and slices' tenants.

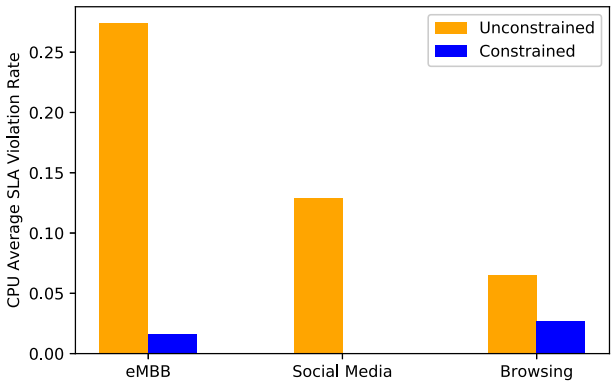


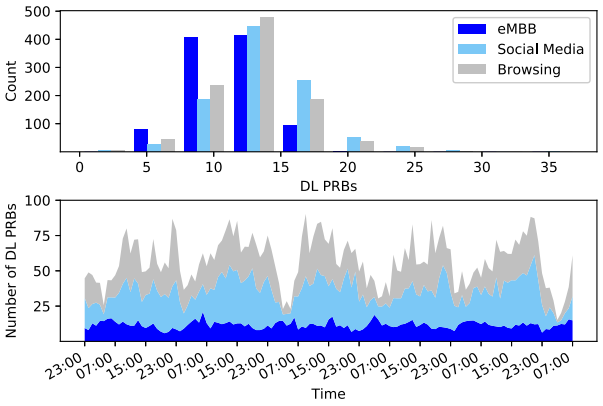
Fig. 6. CPU load average violation rates for constrained SFL vs. unconstrained FedAvg using test dataset with $\alpha = [0, 0, 0]$, $\beta = [4, 7, 10]$ % and $\gamma = [0.01, 0.01, 0.01]$.

C. Performance of Q-Th Percentile SLA

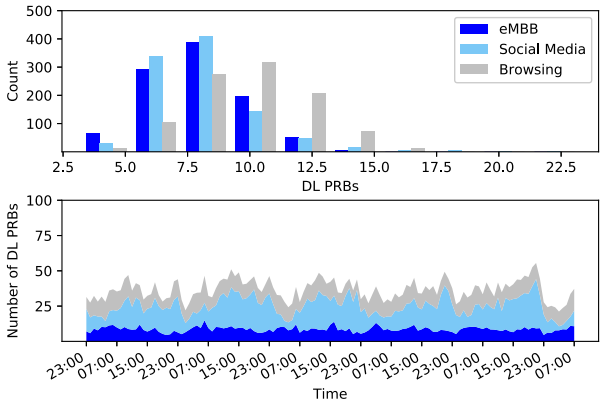
To guarantee slices isolation, the physical operator can control the occupied resources per slice by setting in the SLA an upper threshold on the long-term Q-th resource percentile. This statistical metric is widely used to measure the performance of live networks in terms of e.g., latency and throughput, and yields a more accurate insight on the distribution of a KPI compared with averages. As shown in Fig. 7-(a), the baseline unconstrained predicted DL PRBs per slice are

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

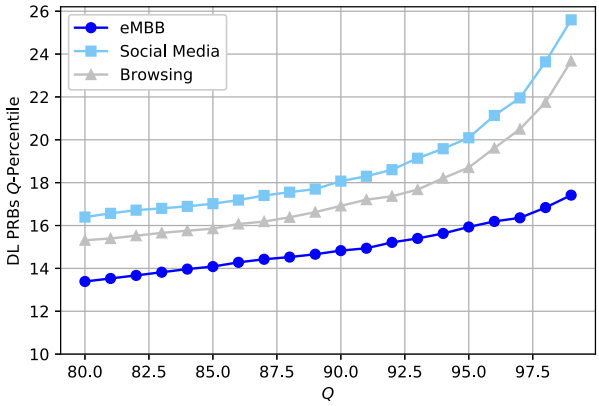
Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه



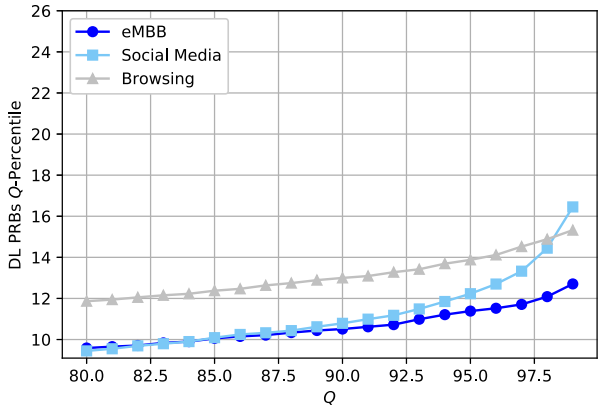
(a) DL PRBs distribution baseline unconstrained FedAvg



(b) DL PRBs distribution with constrained SFL



(c) DL PRBs evolution with baseline unconstrained FedAvg



(d) DL PRBs evolution with constrained SFL

Fig. 7. DL PRBs with 99-th percentile SLA using test dataset for $\pi = [12, 15, 15]$ PRBs.

following the traffic trends but fail to respect the upper bounds π imposed to the 99-th percentile and, as depicted in Fig. 7-(c), reach 26, 24 and 16 PRBs for eMBB, Social Media and Browsing slices, respectively. This difference between slices emanates essentially from their traffic behavior that is the main factor in the unconstrained scenario. In contrast, thanks to the proxy-Lagrangian approach, the constrained SFL model enables the achievement of the dynamic resource allocation of Fig. 7-(b) that presents alternating busy and quiet times, while also enforcing the upper bound on the 99-th percentiles that reach only 12.3, 16.2 and 14.6 for eMBB, Social Media and Browsing slices, respectively, while their curves present also shallow slopes compared to the relatively steep slopes of the baseline unconstrained case. Since the Q-th percentile metric is an increasing function of Q by definition, all the percentiles with $Q < 99$ are also constrained successfully. Therefore, thanks to the introduced constrained federated learning framework, local AEs are able to learn long-term statistical patterns, without sharing their datasets, and adapt their dynamic resource prediction to fulfill metrics defined over an observation window such as the Q-th percentile. The challenge of this learning task is that the constraints are empirically defined over the dataset distribution. Note that for some KPIs, such as throughput, to guarantee a minimum level of service, lower values of Q, e.g., the 1%-percentile, are usually used.

D. Performance of Maximum/Minimum SLA

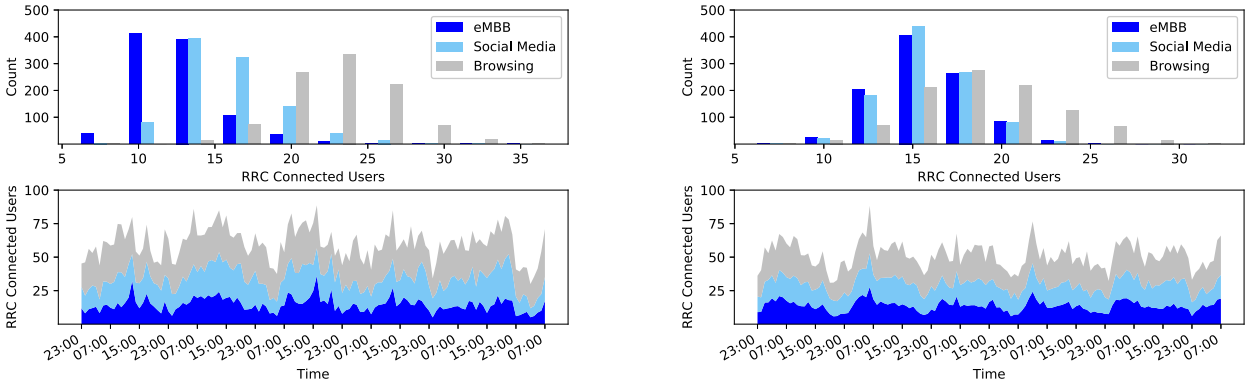
Similarly to practical metrics such as the maximum bit rate (MBR), this type of SLA aims at ensuring maximum and minimum limits for resource allocation. As shown in Fig. 7, the considered resource is the number of RRC connected users licenses wherefore the lower and upper bounds are $\alpha = [10, 0, 0]$ and $\beta = [25, 20, 30]$. In this respect, the histogram distribution of Fig. 8-(a) shows that, for eMBB slice, the minimum is around 7 while for Browsing slice, the maximum is around 33. These two values are not respecting the SLA bounds. By using a constrained model, we statistically enforce the maximum and minimum over the observation period of Fig. 8-(b) that spans 5 days. Specifically, The minimum of eMBB slice is close to 10 and the maximum of Browsing slice is lower than 30 as depicted in the histogram distribution, which is in lone with the imposed bounds.

E. Convergence Time Vs. Overhead

To highlight the trade-offs of the proposed statistical federated learning (SFL), we conduct extensive experiments where we consider an additional baseline, namely, a centralized constrained learning (CCL) model that is trained on the full dataset composed of the aggregation of the K datasets. The training is done using batches of the same size as the local datasets, i.e., 1000 samples. This means that a communication

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com



(a) RRC Users with baseline unconstrained FedAvg

(b) RRC Users with constrained SFL

Fig. 8. RRC Users with maximum/minimum SLA using test dataset, where $\alpha = [10, 0, 0]$ and $\beta = [25, 20, 30]$.

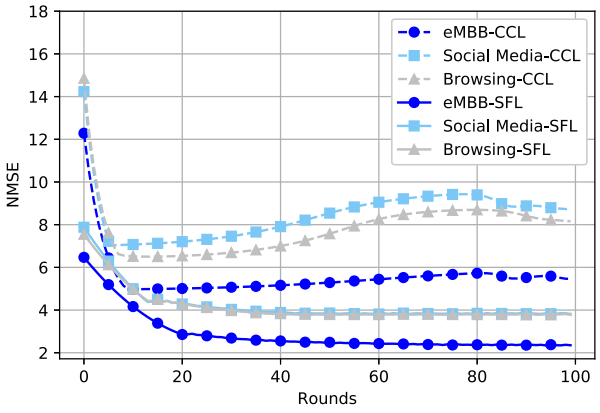


Fig. 9. Convergence of SFL vs. CCL scheme for CDF SLA with $\alpha = [0, 0, 0]$, $\beta = [15, 10, 10]$ PRBs and $\gamma = [0.01, 0.01, 0.01]$.

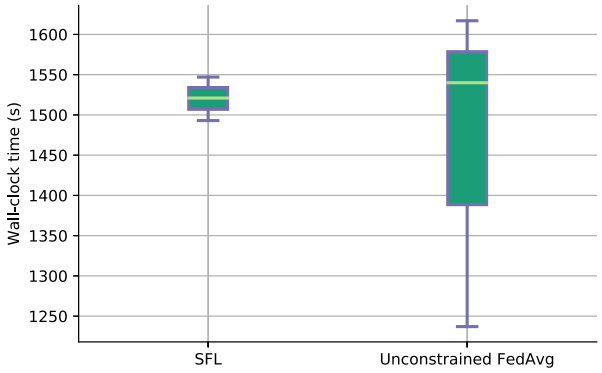


Fig. 10. Wall-clock time of SFL vs. unconstrained FedAvg under the settings of Figure 3 for 100 rounds and on a machine with Core i7 2.20 GHz CPU and 32 GB RAM.

round in the federated setup is equivalent to 100 epochs over a batch in the centralized one. From Fig. 9, we remark that the proposed SFL scheme requires only 5 communication rounds to achieve a similar loss as the centralized scheme. By considering more rounds, the SFL models of the three slices reach lower loss values compared to CCL. This is justified by the fact the fact the AEs take K parallel gradient steps over their datasets compared to a single step in CCL. Besides, the slight increase of loss in CCL after the initial rounds is due to the two-player non-zero-sum game between minimizing the loss and fulfilling the constraints. This behavior is not perceived in SFL due to model averaging. On the other hand, to provide an idea of the order of magnitude for the training time, Fig. 10 depicts the wall-clock time distribution of both SFL and unconstrained FedAvg for 100 rounds and under the same settings of Figure 3, where it turns out that they present approximately a similar time complexity, with FedAvg achieving a faster behaviour in some runs but without being able to enforce the constraints as done by the proposed SFL. Note that this time complexity corresponds to a non-parallel for loop over the 200 CUs participating in the FL training and, therefore, it can be dramatically reduced once the solution is integrated in a real network using a parallel computation framework. Besides, Table IV shows the overhead induced by both CCL and SFL where the samples have been coded in

TABLE IV
OVERHEAD COMPARISON

	Overhead (KB)			
Rounds	50	60	70	80
CCL	18750			
SFL	1055	1266	1477	1688
Gain	$\times$**17.8**	$\times$**14.8**	$\times$**12.7**	$\times$**11.1**

32 bits. Starting from the convergence point of SFL, i.e., round 50, we can achieve more than 10 times overhead gain at the expense of the communication delay. Therefore, SFL turns out to be a more efficient scheme especially when the transmission latency is comparable to the CCL processing delay.

VI. CONCLUSION

In this paper, we have first introduced constrained statistical federated learning (SFL) slice-level resource provisioning models that are able to learn over a data distribution in an offline mode while respecting preset local long-term SLA constraints. Specifically, we have considered three statistical metrics, namely, resource CDF, resource Q-th percentile and maximum/minimum resource bounds, and solved the underlying optimization task using a proxy-Lagangian two player game strategy. We have then shown that the proposed decentralized resource allocation scheme allows for SLA respect

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.

Downloaded from https://iranpaper.ir
http://pooyamoshaver.com
Puyamoshaver@gmail.com
مشاوره انجام پروپزال و پایان نامه

compared to unconstrained FedAvg while also dramatically reducing the communication overhead compared to a centralized constrained deep learning which paves the way for a scalable and sustainable network slicing. By invoking reliability theory, we have finally provided a closed-form analysis for the lower bound of the reliable convergence probability of the proposed federated learning algorithm, where both the respect of SLA and convergence rate are jointly characterized.

REFERENCES

[1] NGMN Alliance. *Description of Network Slicing Concept*. Accessed: Mar. 2019. [Online]. Available: https://www.ngmn.org

[2] C. Marquez *et al.*, "How should I slice my network? A multi-service empirical evaluation of resource sharing efficiency," in *Proc. MobiCom*, 2008, pp. 77–84.

[3] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[4] M. A. Habibi *et al.*, "The structure of service level agreement of slice-based 5G network," Sep. 2018, *arXiv:1806.10426*. Accessed: Sep. 2, 2021. [Online]. Available: https://arxiv.org/abs/1806.10426

[5] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, Jun. 2020. [Online]. Available: https://arxiv.org/abs/1908.06847

[6] *Zero-Touch Network and Service Management (ZSM); Reference Architecture*, document ETSI GS ZSM 002, Aug. 2019.

[7] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*. [Online]. Available: http://arxiv.org/abs/1912.04977

[8] Y. Li, A. Huang, Y. Xiao, X. Ge, S. Sun, and H.-C. Chao, "Federated orchestration for network slicing of bandwidth and computational resource," 2020, *arXiv:2002.02451*. [Online]. Available: http://arxiv.org/abs/2002.02451

[9] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 1205–1221, Jun. 2019.

[10] Z. Yang, M. Chen, W. Saad, C. Seon Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," 2019, *arXiv:1911.02417*. [Online]. Available: http://arxiv.org/abs/1911.02417

[11] F. Fossati, S. Moretti, and S. Secci, "Multi-resource allocation for network slicing under service level agreements," in *Proc. 10th Int. Conf. Netw. Future (NoF)*, Rome, Italy, Oct. 2019, pp. 48–53.

[12] H. Chergui and C. Verikoukis, "Offline SLA-constrained deep learning for 5G networks reliable and dynamic end-to-end slicing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 350–360, Feb. 2020.

[13] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Cognitive network management in sliced 5G networks with deep learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Paris, France, Apr. 2019, pp. 280–288.

[14] J. Zheng, G. de Veciana, and A. Banchs, "Constrained network slicing games: Achieving service guarantees and network efficiency," 2020, *arXiv:2001.01402*. [Online]. Available: http://arxiv.org/abs/2001.01402

[15] B. Khodapanah, A. Awada, I. Viering, D. Oehmann, M. Simsek, and G. P. Fettweis, "Fulfillment of service level agreements via slice-aware radio resource management in 5G networks," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Porto, Portugal, Jun. 2018, pp. 1–6.

[16] H.-B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1273–1282.

[17] J. A. K. Suykens *et al.*, *Advances in Learning Theory: Methods, Models and Applications*. Amsterdam, The Netherlands: IOS Press, May 2003.

[18] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optim. Eng.*, vol. 17, no. 6, pp. 263–287, 2016.

[19] A. Cotter *et al.*, "Training well-generalizing classifiers for fairness metrics and other data-dependent constraints," 2018, *arXiv:1807.00028*. [Online]. Available: http://arxiv.org/abs/1807.00028

[20] A. Cotter, H. Jiang, and K. Sridharan, "Two-player games for efficient non-convex constrained optimization," 2018, *arXiv:1804.06500*. [Online]. Available: http://arxiv.org/abs/1804.06500

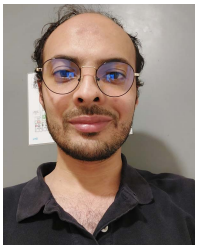
[21] G. J. Gordon, A. Greenwald, and C. Marks, "No-regret learning in convex games," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 360–367.

[22] A. Cotter. *Constrained Optimization (TFCO)*. Accessed: Sep. 2, 2021. [Online]. Available: https://github.com/google-research/tensorflow_constrained_optimization

[23] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*. 5th ed. Hoboken, NJ, USA: Wiley, 2019.

[24] F. Nielsen and K. Sun, "Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures using piecewise log-sum-exp inequalities," 2016, *arXiv:1606.05850*. [Online]. Available: http://arxiv.org/abs/1606.05850

[25] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.



Hatim Chergui (Member, IEEE) received the bachelor's degree in telecommunications engineering from the Institut National des Postes et Télécommunications (INPT), Rabat, Morocco, in 2007, and the Ph.D. degree *(summa cum laude)* in electrical engineering and telecommunications from IMT Atlantique (Télécom Bretagne), Brest, France, in 2015. Since 2008, he has been a Radio Network Planning and Optimization Engineer with extensive industry experience in providing 3G/4G consulting at Huawei Technologies, Morocco. He has worked as a Radio Technologies Expert at Moroccan Operator INWI, Casablanca, Morocco. He is currently the Project Manager of the H2020 MonB5G European Project and a Post-Doctoral Researcher at the Catalan Telecommunications Technology Center (CTTC), Barcelona, Spain. His research interests lie in the area of performance analysis and artificial intelligence for wireless communications. He was a recipient of the IEEE ICC 2020 Best Paper Award. He is an Associate Editor at IEEE NETWORKING LETTERS.

Luis Blanco (Member, IEEE) received the M.Sc. degree in telecommunications engineering, the M.Sc. degree in research on information and communications technologies (MERIT), and the Ph.D. degree in telecommunications engineering from the Polytechnic University of Catalonia (UPC), Spain, in 2006, 2012, and 2017, respectively, the degree in data science and big data from the University of Barcelona (UB), and the master's degree in quantitative techniques for financial markets from UPC. In 2007, he joined the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), where he currently holds a position as a Researcher. He has actively participated in more than twenty research projects funded by public and private entities, including EC-funded projects (MONB5G, SEMANTIC, EMPhAtic, BeFEMTO, and COOPCOM), national projects (5G-TRIDENT, FBMC-SILAN, and GRE3N), networks of excellence in research (SATNEX IV) with the European Space Agency (ESA), and a plethora of industrial projects with international companies (Huawei, Keysight Technologies, Inmarsat, ZIV, and Hispasat). His current research interests include AI/ML for sustainable B5G/6G wireless networks, M2M/the IoT over satellite systems and massive M2M communications for the Internet of Things, and dense wireless networks. He was a recipient of the INNOVATIA Award from the Instituto de Directivos de Empresa (IDE-CESEM), Madrid.

Christos Verikoukis (Senior Member, IEEE) received the Ph.D. degree from UPC, Barcelona, Spain, in 2000. He is currently a Research Director (R4) with the Telecommunications Technological Centre of Catalonia (CTTC/CERCA), Castelldefels, Spain, and an Adjunct Professor with the University of Barcelona. He has authored more than 143 journal articles, over 220 conference articles, three books, 14 book chapters, and five patents. He has participated in more than 30 competitive research projects, while he has supervised 19 Ph.D. students and ten post-doctoral researchers. He is the IEEE ComSoc EMEA Director and a Member-at-Large of IEEE ComSoc GITC. He received the Best Paper Award at the IEEE ICC 2011 and 2020, the IEEE GLOBECOM 2014 and 2015, and the EuCNC 2016, and the EURASIP 2013 Best Paper Award of the Journal on Advances in Signal Processing.

Authorized licensed use limited to: ULAKBIM UASL - Hacettepe Universitesi. Downloaded on March 31,2022 at 07:46:21 UTC from IEEE Xplore. Restrictions apply.