A Softwarized Service Infrastructure for the Dynamic Orchestration of IT Resources in 5G Deployments

Arcangela Rago*†, *Student Member, IEEE,* Giuseppe Piro*†, *Member, IEEE*, Gennaro Boggia*†, *Senior Member, IEEE*, Paolo Dini‡

** Dept. of Electrical and Information Engineering, Politecnico di Bari, Italy*

e-mail: {arcangela.rago, giuseppe.piro, gennaro.boggia}@poliba.it

† CNIT, Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Italy

‡ CTTC, Centre Tecnològic de Telecomunicacions de Catalunya, Spain

e-mail: paolo.dini@cttc.es

ABSTRACT

Thanks to the 5G, telco operators can offer a new set of advanced services to mobile users which makes use of heterogeneous IT resources deployed at the edge of the network. However, their optimal management is not a simple task to accomplish because of the extreme variability characterizing offered services, traffic profile, user distributions, bandwidth, computing, and memory capabilities available for nodes hosting IT resources. To provide preliminary answers in this direction, this paper presents a high-level description of a softwarized service infrastructure, based on the ETSI-NFV specifications, able to dynamically orchestrate IT resources. Specifically, the number of users attached to the base stations and the capabilities of nodes hosting IT resources are continuously monitored through Software-Defined Networking facilities and reported to a high-level orchestrator. Here, a Convolutional Long Short-Term Memory scheme is firstly adopted to provide a spatio-temporal prediction of user distributions and related traffic demands. Then, an optimization problem is executed for configuring location, settings, amount, and usage of IT resources, based on the prediction outcomes. The behavior of the prediction process is deeply investigated. The optimization problem, instead, is described in its preliminary formulation, which gives a clear idea of future research activities in this direction.

Keywords: Software-Defined Networking, ETSI-NFV, Network Optimization, Deep Learning.

1. INTRODUCTION

The integration of Software-Defined Networking (SDN) and Network Function Virtualization (NFV) technologies offers concrete opportunities to develop the fifth generation (5G) of mobile networks, ensuring high performance in terms of scalability and rapid time to market. In this context, telco operators can offer a new set of advanced services to mobile users, which makes use of heterogeneous IT resources deployed at the edge of the network, while increasing the network programmability and reducing CAPital EXpenditure (CAPEX) and OPerational EXpenses (OPEX). In this context, however, the optimal management of IT resources represents a very challenging task to accomplish because of the extreme variability characterizing offered services, traffic profile, user distributions, and the availability of bandwidth, computing, and memory capabilities for nodes hosting IT resources [1], [2].

In the scientific literature, resource management and allocation tasks (i.e., computation offloading, joint communication and computing resource allocation problem) have been addressed through various optimization algorithms focusing on different objectives, ranging from the minimization of the energy consumption or the delay to the maximization of the quality of service [3]. Most of these contributions takes decisions by only considering the current knowledge of the network configuration and conditions. Nevertheless, network optimization can be improved by introducing deep learning techniques, properly tailored to anticipate traffic and mobility patterns [4], [5]. Prediction tasks are generally investigated separately in the current state of the art. Traffic forecasting has been achieved through Convolutional Neural Networks (CNNs), Long Short-Term Memories (LSTMs), or a combination of them [6]. Mobility has been predicted through Markov Chains, Markov Decision Processes, Hidden Markov Models, Bayesian Networks, or Neural Networks for trajectory and location prediction, or a combination of Neural Networks and Bayesian Networks for the prediction of user distributions [7]. A first attempt that jointly performs traffic and mobility prediction by means of a multivariate LSTM architecture is presented in [8]. Nevertheless, at the time of this writing, no contributions consider the usage of this promising instruments for the optimal management and allocation of IT resources in 5G deployments.

Starting from a preliminary contribution presented in [9] by the same authors of this work, this paper extends the current state of the art by proposing a softwarized service infrastructure, based on the ETSI-NFV specifications, which dynamically configures and orchestrates IT resources available in 5G deployments. Specifically, the number of users attached to base stations and the capabilities of nodes hosting IT resources are continuously monitored through SDN facilities and reported to a high-level orchestrator. The Convolutional Long Short-Term Memory (ConvLSTM) scheme is firstly adopted to provide a spatio-temporal prediction of user distributions and related traffic demands. Then, an optimization problem is executed for configuring location, settings, amount, and usage

of IT resources, based on the prediction outcomes. This proposal can be safely implemented also in a large-scale scenario where base stations and network routers are managed by a multi-layer controller structure. In this case, SDN controllers can communicate with a parent controller, before reporting the number of users attached to base stations and the capabilities of nodes hosting IT resources to the orchestrator.

The remainder of the paper is as follows. Section 2 illustrates the proposed service infrastructure. Section 3 presents the behavior of the prediction process by considering three possible use cases. Section 4 provides some guidelines for the design of a network optimization algorithm implemented by the orchestrator. Finally, Section 5 concludes the paper and draws future research activities.

2. THE PROPOSED SERVICE INFRASTRUCTURE

The softwarized service infrastructure presented herein aims at dynamically orchestrating IT resources in 5G deployments, supporting a wide range of applications. E-Health, autonomous driving, and augmented/virtual reality are possible examples of applications with different communication and computing requirements (see Table I). Note that the latency requirement, ranging from 1 ms to 100 ms [10], is a highly critical aspect to be taken into account for the design of the orchestration framework.

Table 1. Use case requirements.

Use case	Latency	Bandwidth	Memory	Computation
e-Health	1-10 ms [10]	100 Mbps [10]	8 GB [11]	1.2 GHz [11]
Autonomous Driving	10-100 ms [10]	700 Mbps [10]	16 GB [12]	2.1 GHz [12]
Virtual Reality	1 ms [10]	1 Gbps [10]	32 GB [13]	3.5 GHz [13]

Fig. 1 shows the proposed service infrastructure. It embraces different domains, whose network equipments are managed by a single SDN controller, namely the domain controller. At the large-scale, the configuration of different domains can be managed through a hierarchical and highly scalable architecture of controllers [10]. Specifically, multiple domain controllers, that are connected to base stations, nodes hosting IT resources, and network routers, are specialized controllers in charge of intra-domain services. They are coordinated by a parent controller or directly orchestrated by the network orchestrator, which interoperates with domain controllers to provide end-to-end and inter-domain services. The system model described herein simply considers base stations and network nodes belonging to a single domain. However, it can be safely extended to describe the aforementioned hierarchical scenario.

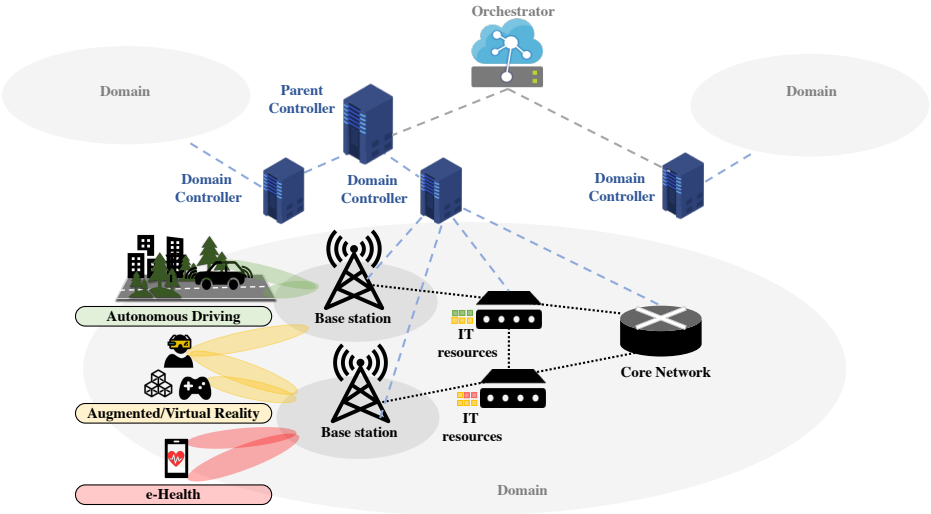


Figure 1. The proposed service infrastructure.

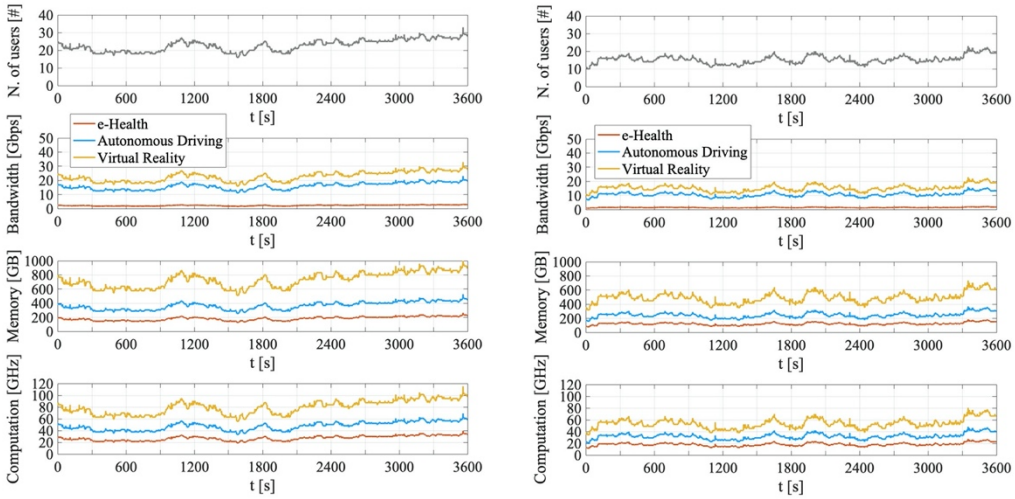
The set of base stations available within the considered domain is denoted with J. At the same time, the set of users attached to the j-th base station is denoted with I_j and the whole set of users I is given by their combination. Based on the targeted application, the i-th user requests a communication bandwidth of b_i and an upper bound of the communication latency equal to τ_i, as well as IT resources with memory and computing capabilities equal to m_i and c_i, respectively. At the edge of the network there are N nodes hosting IT resources. Given the n-th node belonging to N, its memory and computing capabilities are identified with $\mathcal{M}_n$ and $\mathcal{C}_n$.

The role of the SDN controller is to periodically monitor the number of users in I_j, attached to each base station, the communication bandwidth available between the j-th base station and the n-th node hosting IT resources, and the computation and memory capabilities available for the n-th node hosting IT resources. This information is delivered to the high-level orchestrator for predicting the user distribution and related traffic demands in future time instants and performing the optimal allocation of IT resources among mobile users. The outcomes of the allocation problem are used to configure the whole softwarized service infrastructure through SDN facilities.

3. PREDICTION OF USER DISTRIBUTIONS AND RELATED TRAFFIC DEMANDS

Spatio-temporal distributions of users are captured by SDN controllers and collected by the orchestrator, which can consequently perform the prediction of user distributions and related traffic demands. Specifically, the prediction architecture adopted for this purpose is based on the ConvLSTM scheme, that exploits LSTM memory cells and the convolutional operation, through which it can extract temporal and spatial correlations of data, respectively [9]. The predicted number of users attached to different base stations is then used to quantify bandwidth, computing, and memory requirements to be taken into the account by the optimization algorithm.

The adopted dataset refers to a one-month distribution of taxi cabs in the center of Rome. The considered geographical area of around 110 km^2 has been divided using 11 × 10 square cells, so that the j-th grid cell, attached to the j-the base station, covers a square area of 1km × 1km. Fig. 2 shows the predicted number of users attached to two reference base stations (i.e., $j = 45$ and $j = 55$). Furthermore, by assuming that each user enjoys, at the same time, e-Health, autonomous driving, and virtual reality services, the aggregate requests of communication and computing resources coming from the two reference base stations are depicted as well.



(a) *(b)*

Figure 2. Predicted number of users and aggregated communication and computing resources estimated for e-Health, autonomous driving, and virtual reality use cases for (a) $j = 45$ and (b) $j = 55$.

4. PRELIMINARY CONSIDERATIONS FOR THE NETWORK OPTIMIZATION

This Section provides a preliminary formulation of the optimization problem that considers the communication latency as the main objective for configuring location, settings, amount, and usage of IT resources. The overall latency l_{ijn} experienced by each user i, connected to the j-th base station and served by the n-th node hosting IT resources, can be considered as the sum of three contributions [10]:

$$l_{ijn} = l_{ij}^{radio} + l_{ijn}^{backhaul} + l_{in}^{execution},$$

where l_{ij}^{radio} is the radio interface component, $l_{ijn}^{backhaul}$ is the backhaul latency between the j-th base station and the n-th node hosting IT resources at the edge of the network, and $l_{in}^{execution}$ is the execution latency required to the specific n-th node hosting IT resources for serving the user i.

The preliminary formulation of the optimization problem executed by the orchestrator is provided in Table 2. The reported considerations will be deeply investigated in future research activities.

Table 2. Considerations related to the optimization problem.

Aspect	Analytical formulation	Description
Possible optimization objective	$\min_{\alpha_{in}} \left\{ \sum_{j \in J} \sum_{i \in \widehat{I}_j} \left[l_{ij}^{radio} + \sum_{n \epsilon N} \alpha_{iz} \left(l_{ijn}^{backhaul} + l_{in}^{execution} \right) \right] \right\}$ where α_{in} is a binary decision variable that is 1 if the user i is served by the n-th node, otherwise it is 0.	Given the predicted number of users attached to each base station $\widehat{I}_j$, the optimization problem could minimize the latencies experienced by users.
Constraints on memory and computation capabilities	$\sum_{i \in I} \alpha_{in}\, m_i \leq \mathcal{M}_n, \forall n \in N$ $\sum_{i \in I} \alpha_{in}\, c_i \leq \mathcal{C}_n, \forall n \in N$	The number of users served by the n-th node cannot exceed its memory and computation capabilities.

Communication latency constraint	$l_{ij}^{radio} + \sum_{n \epsilon N} \alpha_{in} \left(l_{ijn}^{backhaul} + l_{in}^{execution} \right) \leq \tau_i, \forall i \in I$	The overall latency experienced by each user i cannot exceed the upper bound of its communication latency τ_i.
Deployment of new IT resources	$\sum_{i \in I} m_i \geq \sum_{n \in N} \mathcal{M}_n$ $\sum_{i \in I} c_i \geq \sum_{n \in N} \mathcal{C}_n$	Available IT resources cannot meet all the user requests. Telco operators must install new IT resources at the edge.

5. CONCLUSIONS

This work has presented a softwarized service infrastructure, based on the ETSI-NFV specifications, able to dynamically configure and orchestrate IT resources. We have performed the spatio-temporal prediction of user distributions and related traffic demands through a Convolutional Long Short-Term Memory scheme. Then, the outcomes of the prediction process have been thoroughly examined by considering e-Health, autonomous driving, and virtual reality use cases. Moreover, the prediction outcomes have been adopted to sketch an optimization problem to minimize the overall latency experienced by mobile users and consequently configure IT resources hosted in edge nodes. Further research activities will operatively implement optimization algorithms, exploiting the prediction process to anticipatory orchestrate the network infrastructure.

ACKNOWLEDGEMENTS

This work was supported by the Apulia Region (Italy) Research project INTENTO (36A49H6) and the PRIN project no. 2017NS9FEY entitled “Realtime Control of 5G Wireless Networks: Taming the Complexity of Future Transmission and Computation Challenges” funded by the Italian MIUR.

REFERENCES

[1] Z. Zaidi, V. Friderikos, Z. Yousaf, S. Fletcher, M. Dohler and H. Aghvami, "Will SDN Be Part of 5G?," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3220-3258, 2018.

[2] R. Alvizu *et al*., "Comprehensive Survey on T-SDN: Software-Defined Networking for Transport Networks," in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2232-2283, 2017.

[3] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, “Machine Learning Meets Computation and Communication Control in Evolving Edge and Cloud: Challenges and Future Perspective,” *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 38–67, 2020.

[4] B. Ma, W. Guo, and J. Zhang, “A Survey of Online Data-Driven Proactive 5G Network Optimisation Using Machine Learning,” *IEEE Access*, vol. 8, pp. 35 606–35 637, 2020

[5] A. R. Mohammed, S. A. Mohammed and S. Shirmohammadi, "Machine Learning and Deep Learning Based Traffic Classification and Prediction in Software Defined Networking," *IEEE International Symposium on Measurements & Networking (M&N)*, Catania, Italy, 2019, pp. 1-6.

[6] C. Zhang, P. Patras, and H. Haddadi, “Deep Learning in Mobile and Wireless Networking: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[7] H. Zhang and L. Dai, “Mobility Prediction: A Survey on State-of-the-art Schemes and Future Applications,” *IEEE Access*, vol. 7, pp. 802–822, 2018.

[8] L. Chen, D. Yang, M. Nogueira, C. Wang, D. Zhang *et al.*, “Data-Driven C-RAN Optimization Exploiting Traffic and Mobility Dynamics of Mobile Users,” *IEEE Transactions on Mobile Computing*, 2020.

[9] A. Rago, P. Ventrella, G. Piro, G. Boggia, and P. Dini, "Towards an Optimal Management of the 5G Cloud-RAN through a Spatio-Temporal Prediction of Users' Dynamics", *Proc. of IEEE Mediterranean Communication and Computer Networking Conference (MedComNet)*, June, 2020.

[10] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, “A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[11] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, “Exploiting Smart e-Health Gateways at the Edge of Healthcare Internet-of-Things: A Fog Computing Approach,” *Future Generation Computer Systems*, vol. 78, pp. 641–658, 2018.

[12] M. Jung, S. A. McKee, C. Sudarshan, C. Dropmann, C. Weis, and N. Wehn, “Driving into the Memory Wall: the Role of Memory for Advanced Driver Assistance Systems and Autonomous Driving,” in *Proc. of International Symposium on Memory Systems*, 2018, pp. 377–386.

[13] R. Albert, A. Patney, D. Luebke, and J. Kim, “Latency Requirements for Foveated Rendering in Virtual Reality,” *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 4, pp. 1–13, 2017.