An AI Framework for Fostering 6G towards Energy Efficiency

Raffaele Bolla
DITEN - University of Genoa - Genoa, Italy
raffaele.bolla@unige.it

Roberto Bruschi
DITEN - University of Genoa - Genoa, Italy
roberto.bruschi@unige.it

Franco Davoli
DITEN - University of Genoa - Genoa, Italy
franco.davoli@unige.it

Lorenzo Ivaldi
DITEN - University of Genoa - Genoa, Italy
lorenzo.ivaldi@unige.it

Chiara Lombardo
CNIT – S2N National Lab - Genoa, Italy
chiara.lombardo@cnit.it

Beatrice Siccardi
CNIT – S2N National Lab - Genoa, Italy
DITEN - University of Genoa - Genoa, Italy
beatrice.siccardi@tnt-lab.unige.it

***Abstract*— The sixth generation of radio mobile networks (6G) is expected to employ Artificial Intelligence (AI) to support this new network evolution. Along with the enabling technologies of its predecessor, namely edge computing and network slicing, AI will realize intelligent communication and networking, but at a high risk of severely increasing the demands on computing resources and the related carbon footprint. In order to avoid such a curse, this paper proposes an AI framework allowing to achieve tight proportionality between the workload produced by vertical applications and network slices, and the energy consumption induced in the infrastructure, and enabling beyond 5G stakeholders to become part of a green business model. The paper outlines a high-level architecture able to support the framework, as well as the role of the algorithms composing it. Results show that the proposed framework allows for energy savings over 100% while fulfilling the application requirements.**

Keywords*— *B5G, Artificial Intelligence, Edge Computing, Green Networking

I. INTRODUCTION

The fifth generation of radio mobile networks (5G) has fostered the rise of a wider range of services, such as large-scale Internet of Things (IoT), by adopting edge computing [1] and network slicing [2] as enabling technologies: customized instances of the 5G network, providing the desired network and radio services, can be assigned on demand to vertical stakeholders. Edge computing closes the loop by hosting vertical applications into the 5G infrastructure, and directly attaching them to the network slice terminations in neighboring geographical facilities. While the vision and requirements for moving Beyond 5G (B5G) are still in their infancy, it is expected that Artificial Intelligence (AI) will represent a game changer to penetrate the whole network and realize the ultimate goal of "pervasive intelligence" [3].

In order to truly impact our society, these enabling technologies would be of little or no use unless the economic and technological access barriers are removed for a massive number of stakeholders, by enabling new business opportunities at affordable costs: only a mass market can assure the revenue levels needed to balance the huge investments from telecom infrastructure and technology providers. However, the widespread diffusion of 5G and the evolution towards B5G will not happen without any costs and risks, which might even undermine the technological foundations and the economic sustainability of this technology. If not properly addressed, the intrinsic distributed and pervasive nature of B5G technologies will cause a noticeable usage and deployment increase of computing resources, of the associated infrastructure Operating Expenditure (OpEx) and Capital Expenditure (CapEx), and, consequently, of their carbon footprint and energy requirements [4], much higher than in today's scenario. Under this perspective, the scalability levels provided by a simple "business-as-usual" evolution beyond 5G and edge computing might not be sustainable and sufficient to support the rise of a new mass-market and might be even constrained by the availability of resources and energy to supply power-hungry edge datacenters.

In order to fasten and foster the sustainable rise of a new mass-market of novel vertical applications with challenging, heterogeneous, and time-varying requirements, an innovative, full value-chain ecosystem is required to make all the B5G stakeholders becoming integral part of a win-win business model, common to the best green economy practices, by promoting behaviors (e.g., consuming resources as-a-Service) that reduce the power consumed in the infrastructure through incentives.

This ambitious goal will be achieved only in the presence of fully automated orchestration tools able to sense relevant Key Performance Indicators (KPIs) and events, decide when to trigger elasticity- and agility-driven (proactive) operations, and efficiently perform them by coordinating all the involved architectural building blocks. As a consequence, the evolution of the 5G Service-Based Architecture (SBA) will see the introduction of new features for supporting of such tools, for example by collecting events and analytics streams from both the infrastructure and the vertical applications.

This paper outlines the most relevant principles to be embraced for the development of a B5G ecosystem. The smart, fast, and automated scaling of vertical application and related slices across the geographically distributed B5G edge-cloud continuum will allow for workload redistribution upon user or infrastructure-driven events and, when paired with the usage of diverse physical resources, will allow for better trade-offs between consumption and latency. Following such principles, the paper sketches the characteristics required for the successful deployment of such an ecosystem. The envisioned architecture stems from the experience acquired in several recent European Projects [5] [6] which provided the technologies required for the automated, intelligent and dynamic management of resources prescribed by the AI [7] [8], and entails the adoption of the ETSI ENI (Experiential

978-88-87237-53-5

Networked Intelligence) specification [9] for defining AI policies fed by the B5G enhancement of the SBA, especially via the Network Data Analytics Function (NWDAF) and the Management Data Analytics Function (MDAF).

An evaluation is carried out to determine the impact and the potential outlets that can be enabled by our approach. The evaluation considers a cell-tower edge datacenter deployment on a metropolitan area, and is based on real, publicly available datasets characterizing the infrastructure and traffic of one of the main Italian mobile operators. Since the currently available data regard 4/5G networks, estimates on the future scenario are applied when needed. Results assess both the consistency of the proposed framework and its effectiveness in bringing huge energy savings for the whole ecosystem and, more importantly, highlight several aspects and best practices to drive the design of an energy aware B5G ecosystem.

The remainder of the paper is organized as follows. Section II presents the vision towards B5G, while Section III focuses on the AI framework. The evaluation is reported in Section IV and conclusions are drawn in Section V.

II. Vision Beyond 5G and Design Pillars

This section introduces the main design pillars at the basis of the vision we propose for moving beyond 5G, namely, *Edge Agility* (Section II.A) and *Green Elasticity* (Section II.B) before presenting a high-level architecture aiming to support the intelligent and dynamic management of resources prescribed by AI, which is outlined in Section II.C.

A. Edge Agility

Edge Agility regards the smart, fast, and automated horizontal scaling of vertical application and related slices across the geographically distributed B5G edge-cloud continuum. It will be a key actuation mean to redistribute the workload according to user or infrastructure-driven events, like user mobility, possible transparent replacement/migration of the workload to meet the availability of renewable energy sources, or energy-driven reconfiguration of radio access networks.

Edge Agility can be roughly seen as a sort of handover procedure from the SBA/application point of view. It will enable the B5G SBA to autonomically proacting/reacting to UE handovers or policy-driven events by triggering joint management operations on slice network functions and on the vertical application, while interacting with the control plane to assure seamless operations through suitable procedures based on cloud-native service-mesh routing. By redistributing the workload, it will be possible to move the latency budget accordingly between connectivity (in terms of PDU session or distance from the edge) and computing (the time taken by software processes composing network functions or application components).

B. Green Elasticity

Green Elasticity refers to the capability of scaling performance by opportunistically trading-off (and mixing) diverse resources, as for example computing resources for low-latency storage or network slices, to enable smart vertical scalability across the three domains of B5G environments: from the vertical domain, through the network (slice), down to the infrastructure.

On one side, hardware acceleration can significantly lower processing latency with respect to pure software artefacts; on the other side, the energy efficiency of hardware acceleration strictly depends on the workload volume offloaded from the software level. Hardware acceleration engines usually exhibit low power-consumption dependency against their usage, and significant energy savings can only be achieved by putting the engines into standby low power modes. Therefore, hardware acceleration becomes energetically/environmentally advantageous against pure software processing when applied to large volumes of the time-varying workloads, or when speeding up one element in the NF/vApp component chain can lead to optimal end-to-end configurations/deployments.

Green Elasticity is a crucial mean for dynamically distributing the time-varying, end-to-end latency budgets of vertical applications across the domains, while holistically optimizing the trade-off between the energy/carbon footprint and the performance of network and application artefacts. Through this paradigm, elasticity operations will be no longer constrained to a single domain, but they will rather propagate to the other domains (i.e., through proper interfaces between the SBA and the Application Function – AF). This will allow jointly adapting/consolidating the service meshes of vertical applications and network slices to optimally exploit the available computing/networking resources.

C. Architectural Vision

Since it is expected that the infrastructure and the (vApp and network) workload will be geographically distributed in 5/6G environments, there is a limited aggregation/multiplexing gain, differently from what is usually happening in cloud computing datacenters. vApps and network slices need to deal with human and device mobility, and with local workload surges on resource-constrained edge datacenters. The architectural vision that we propose for supporting such environments is depicted in Figure 1, and includes the vertical application domain, the network platform domain and the B5G network infrastructure. This breakdown identifies the main involved stakeholders as well, in accordance with [10].

Since the overall ecosystem should target a fully automated control of all the resources/services at any layer to allow the application lifecycle management in a "multi-tenant" (i.e., hosting multiple overlaying systems) and "multi-domain" (i.e., exploiting the resources from multiple underlying systems) environment, vertical industries are expected to orchestrate and manage their applications through a Vertical Application Orchestrator (VAO), while the interconnection of application components running in different datacenters towards UEs and among themselves should be provided by the NFV Orchestrator (NFVO) in the network platform. Isolation in different tenant spaces of vertical applications and NFV services is required to work on Virtual Infrastructure Managers (VIMs)-level partitions, distributed across the Wide-Area Network (WAN), without competition on shared network and computing resources. Moreover, vApps and network slices should be realized as meshes of "stateless" microservices equipped with sidecars to handle reconfiguration, adaptation, and fast deployment operations in a flexible and efficient way.

In order to ensure the success and applicability of any solutions, it is essential to rely on the frameworks and trend considered by standard making organizations (e.g., 3GPP and ETSI). Our proposed architecture is no exception: the goal is, on the one hand, to extend recent specifications of 3GPP Rel. 16 and, on the other hand, to rely on the ETSI ENI

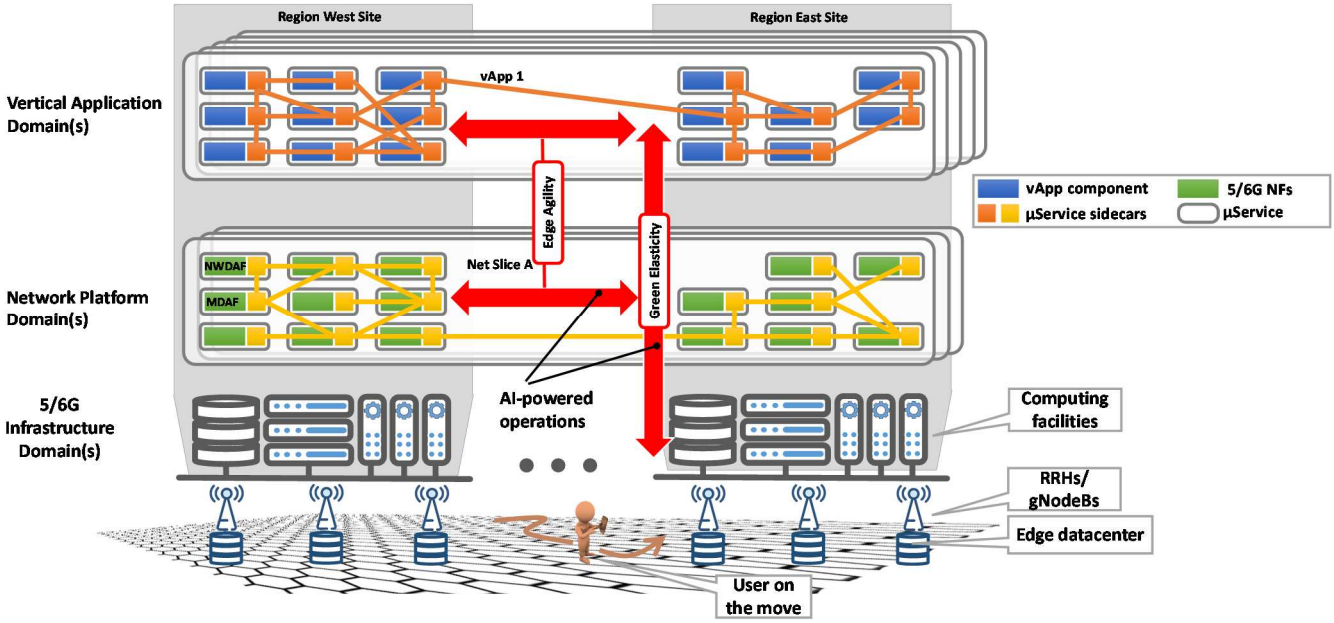

Figure 1. The proposed vision, including the main stakeholders and the design pillars.

specification to adjust offered services through AI and context-aware policies. In more details, the B5G NWDAF and MDAF will provide the observability of cross-domain metrics related to the events and analytics streams coming from the infrastructure and from the vertical applications, and the collected data will feed the AI engines natively integrated in the ENI framework to drive the automated, intelligent and dynamic management of resources.

III. Exploiting AI for Energy Efficient B5G Environments

In order to avoid a waste of resources and the related increase of the carbon footprint, AI approaches will be of paramount importance to leverage on the information on OpEx- and energy-aware metrics extracted from the infrastructure and relevant events or data, to enable proper and cognitive decisions/operations for each stakeholder and become a part of the optimization control loop. Reinforcement learning mechanisms will be used to dynamically analyze the performance of network slices against the intents of vertical applications, and to perform proactive lifecycle operations or to provide feedback information/incentives to vertical stakeholders for reducing their resource usage footprint and induced energy consumption.

The architecture presented in Section II allows to properly "divide and conquer" the complex and multifaceted decision processes, and to make the different architectural building blocks cooperating through AI-based swarm-intelligence and multi-agent system schemes. In more details, the NWDAF will profile UE mobility in a per-slice and per-application fashion, while the MDAF will collect resource usage statistics of network functions and vertical applications' components. AI-based engines will use such information to infer the obtained patterns with the distribution of run-time data and software images to automatically select strategies to minimize the impact on the infrastructure and meet the application intents and network requirements. This will allow closed-loop reactions to be faster and more scalable, since particular events, information flows, or aspects will not be explicitly propagated to all the control layers.

Since different operations performed upon the same trigger event might have a different impact on the infrastructure and on the induced resource usage footprint, the approach proposed in this paper consists of centralizing decisions based on distributed engines, as described in the following section.

A. *Example of an AI Framework*

The general considerations outlined in the previous section can be translated into countless solutions. In this section, without entering the details on the actual algorithms, we present a potential approach to address an AI-driven energy efficiency framework for the upcoming B5G ecosystem. This approach stems from the experience gained in several past European projects [5], [12] combined with AI features, and consists of applying a hierarchy of control policies to different architecture layers/stakeholders to promote overall energy efficiency. Namely, it is composed of three stages:

1. Edge-level policy: each edge datacenter keeps track of its activity and resource usage and exploits trending patterns to apply power saving strategies [13] accordingly. The data monitored by the NWDAF and the MDAF is routinely communicated to the WAN-level policy that aggregates it in a privacy-conscious fashion.
2. WAN-level policy: this policy uses the data coming from the edge datacenters to lead the deployment and orchestration of vertical applications and NFV services in ways that improve energy consumption while respecting performance constraints. The WAN-level policy merges these data with the SLAs of the deployed services (from the AF) and with the WAN topology to obtain the highest energy savings allowed by the admissible latency.
3. Green economy policy: in order to make the B5G stakeholders acting on virtual layers more aware of power consumption, one possible solution would be to (economically) incentivize them to adapt their operations in a way that can be more efficiently handled by the underlying infrastructure. This is a very complex

task that involves, at least, mapping the energy consumption ascribable to the servers' and network elements' hardware components to the execution environments running on top of them, cooperating with the VAOs and NFVO to estimate the resource usage footprint of B5G applications and NFV services, and producing green business models to promote energy-conscious behaviors.

Stage 1 is performed independently by the edge datacenters, while 2 and 3 are in charge of a centralized ENI engine. In the next section we delve into the first two stages and try to draw some quantitative considerations on the effectiveness of our proposed framework.

IV. Evaluation

This section presents several results assessing the soundness of our proposal. Namely, Section IV.A focuses on the policy applied at the individual edge datacenters, while Section IV.B extends the AI framework to the WAN, and further considerations are collected in Section IV.C.

Since real traffic traces of mobile users are not freely available, we relied on the Internet traffic activity obtained by the Telecom Italia Mobile (TIM) Open Big Data initiative [14]. The "Milano Grid" dataset reports the level of interaction of the users with the mobile phone network over a reference metropolitan area that includes Milan, Italy, and neighboring cities, covering an area of 552.25 km^2. We translated such levels into the traffic rate of a sample application belonging to the 3GPP Discrete Automation use case [15] (delay budget corresponding to 10 ms) by scaling them proportionally to the Cisco Mobile Visual Networking Index (VNI) mobile speed forecasts [16].

Since co-location to exploit readily available sites has been proved to reduce deployment costs and improve Power Usage Effectiveness (PUE) [17], for the edge datacenters, we considered a deployment at tower sites, composed of less than ten Dell Precision T5810 workstations as servers, with Intel Xeon CPU E5-1650, 12-core at 3.5 GHz, and 32 GB RAM.

Finally, as the installation of gNodeBs in the area is still at a preliminary stage, we took into account the distribution of the LTE eNodeBs deployed by TIM over the reference area, as reported in the OpenCellID database [18] and depicted in Figure 2.

A. Edge-Level Policy

We considered the traffic activity of a workweek, from 4th to 8th November. Each edge datacenter is in charge of monitoring its incoming traffic in order, on the one hand, to determine useful estimates for adapting its capacity and, on the other hand, to provide aggregated, anonymized data to the centralized AI mechanisms. For example, Figure 3 reports the occupation for one of the edge datacenters over the course of five days. As it can be seen, the day-night traffic profile is quite constant over the working days, so data from previous days can provide the forecast for the following ones and allow applying energy saving techniques to the datacenters. As shown in Figure 4, the error of the estimated system occupation against the measured one is below 1%.

The edge datacenter can perform these operations independently, in order to offload the central AI and guarantee a higher level of security and privacy to the verticals deploying applications instances in their facility. At the same time, the provided data contributes to the overall energy savings as shown in the following section.

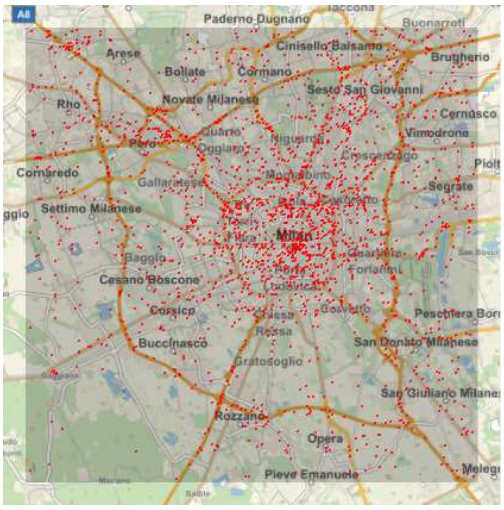

Figure 2. Antennas distributed over the reference area.

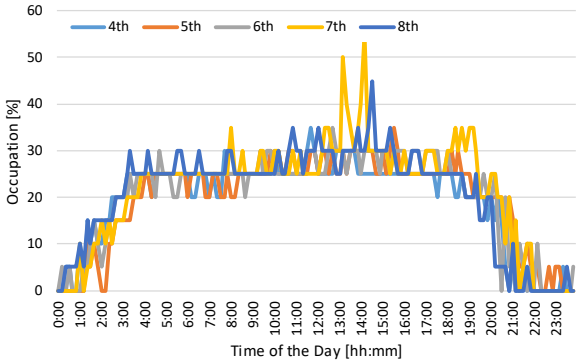

Figure 3. Occupation of one edge datacenter throughout a working week.

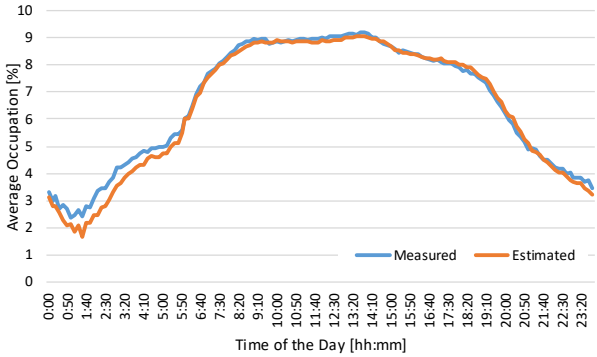

Figure 4. Average measured occupation of the whole system and average occupation estimated over a working week.

B. WAN-Level Policy

By solely applying the edge-level policy, the overall power consumption is already drastically reduced. However, even more ambitious goals can be achieved by pairing this mechanism with a centralized AI allowing to involve all the key players in being part of the energy saving target.

In this respect, by applying the AI framework, it is possible to identify a sub-set of edge datacenters allowing to both satisfy the SLAs and to reduce the overall power consumption. By encouraging verticals to deploy their applications onto such datacenters, energy efficiency is shared across all stakeholders.

As expected, by sharing the overall traffic among a subset of datacenters and so leaving most of the servers in the system powered off, the overall power consumption is drastically reduced. Figure 5 shows the average power consumed on November 8th by applying the WAN-level policy against the one at the edge-level alone. Savings are over 100% throughout the whole day-night profile. Moreover, since the AI algorithm is designed to use latency as a constraint in the selection of the subset of datacenters, the average delay depicted in Figure 6 shows a performance decay with respect to the edge-level policy alone, but the figure still satisfies the application requirements.

C. *Discussion and Future Work*

The proposed results have been presented with the purpose of identifying the applicability and effectiveness of the proposed AI framework, as well as to better identify room for improvement and potential drawbacks.

At the current stage of the 3GPP Release 16, the role of the NWDAF and MDAF functions has been defined but their actual implementation is far from being fully specified. The authors believe that this specification will be one of the main enablers both for the success of 5G technologies and for their evolution B5G.

In Section III, we mentioned the relevance of green economy policies aimed at incentivizing cooperation among stakeholders to make energy savings a common goal. While these policies can be seen as a future work, it is still worth providing some preliminary considerations.

One of the main problems is that the decisions taken at the verticals' level are the main drivers of the power consumption, but the impact actually affects the OpEx of the infrastructure provider who owns the hardware.

In order to quantify the consumption ascribable to the other stakeholders involved in the ecosystem, a set of cross-domain observability mechanisms and analytics are required to evaluate the energy and the carbon footprint that a vertical application, a slice, or the overall B5G network is inducing onto the edge-cloud infrastructure. To reach a holistic green ecosystem and to make all the stakeholders aware of the footprint they induce, specific mechanisms must be put in place to suitably process, infer, and expose this information at both the B5G SBA and vertical application (and their network slices) levels. Hence, the SBA needs to be extended to acquire these energy consumption metrics, and to further classify and expose them to the accountable vertical. Moreover, adaptive AI-driven analytics must be developed to collect hardware-level energy consumption metrics, divide/map them onto each hosted tenant and expose them to the accountable vertical.

V. Conclusions

With 5G networks finally being deployed worldwide, industry and academia are starting to shift their focus to B5G to outline the main requirements and upcoming features for the next generation of mobile communications.

While edge computing has been introduced by design already in 5G, it will play an even more crucial role in the presence of artificial intelligence (AI), which will pervade the infrastructures to realize intelligent communication and networking. However, enabling the dynamic distribution of traffic and resources across the geographical distributed network edge will cause a noticeable increase of computing

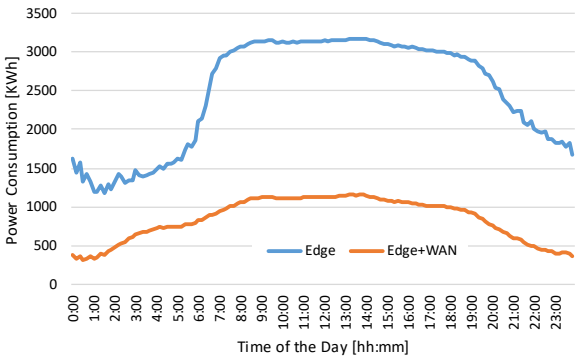

Figure 5. Average power consumption of the system on November 8th obtained by applying the only the edge-level policy alone and along with the WAN-level policy to the ecosystem.

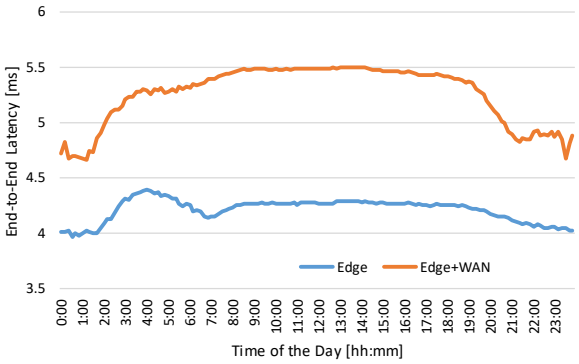

Figure 6. Average latency of the system on November 8th obtained by applying the only the edge-level policy alone and along with the WAN-level policy to the ecosystem.

resources and of the related carbon footprint, unless all the involved stakeholders become part of a full value-chain ecosystem and actively contribute to its energy efficiency.

In this respect, this paper has sketched an architectural vision to support such an ecosystem, and it has proposed an AI framework to allow all the stakeholders jointly targeting energy efficiency. By propagating information collected at the various layers, from the physical infrastructure to the applications, in aggregated forms that guarantee privacy of the users' data, the framework enables proportionality between the workload produced by vertical applications and network slices, and the energy consumption induced in the infrastructure. A green business model (out of the scope of this paper) providing incentives to promote such behaviors would close the energy efficiency loop for all the involved players.

The proposed evaluation has been presented with the purpose of identifying the applicability and effectiveness of the proposed AI framework, as well as to better understand room for improvement and potential drawbacks. Results show that the framework, that interacts with both the verticals and the other Telco-level blocks, allow for savings over 100% while fulfilling the application requirements.

Acknowledgment

This work has been supported by the Horizon 2020 5G-PPP Innovation Action 5G-INDUCE (Grant Agreement no. 101016941).

References

[1] ETSI GS MEC 002, "Mobile Edge Computing (MEC); Technical Requirements", version 1.1.1, March 2016. URL:

http://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/01.01.01_60/gs_MEC002v010101p.pdf.

[2] X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina, "Network Slicing in 5G: Survey and Challenges," IEEE Commun. Mag., vol. 55, no. 5, pp.94-100, May 2017.

[3] G. Liu et al., "Vision, requirements and network architecture of 6G mobile network beyond 2030," in China Communications, vol. 17, no. 9, pp. 92-104, Sept. 2020, doi: 10.23919/JCC.2020.09.008.

[4] "Between 10 and 20% of electricity consumption from the ICT* sector in 2030?", https://www.enerdata.net/publications/executive-briefing/between-10-and-20-electricity-consumption-ict-sector-2030.html.

[5] MATILDA - A Holistic, Innovative Framework for Design, Development and Orchestration of 5G-ready Applications and Network Services over Sliced Programmable Infrastructure, http://www.matilda-5g.eu/.

[6] 5G-INDUCE - Open cooperative 5G experimentation platforms for the industrial sector NetApps, https://www.5g-induce.eu/.

[7] R. Bolla, R. Bruschi, F. Davoli, C. Lombardo, J. F. Pajo, "Multi-site Resource Allocation in a QoS-Aware 5G Infrastructure," IEEE Transactions on Network and Service Management, doi: 10.1109/TNSM.2022.3151468.

[8] R. Bruschi, F. Davoli, C. Lombardo, J. F. Pajo, "Managing 5G Network Slicing and Edge Computing with the MATILDA Telecom Layer Platform", Computer Networks, vol. 194, pp. 1-14, April 2021.

[9] ETSI Experientaial Networked Intelligence (ENI) home page at https://www.etsi.org/committee/eni?tmpl=component.

[10] 3GPP, "Study on Management and Orchestration of Network Slicing for Next Generation Network," TR 28.801, version 15.1.0, Jan. 2018.

[11] ETSI Experientaial Networked Intelligence (ENI) home page at https://www.etsi.org/committee/eni?tmpl=component.

[12] The low Energy COnsumption NETworks (ECONET) Project. https://www.econet-project.eu/.

[13] R. Bolla, R. Bruschi, A. Carrega, F. Davoli, "Green Net. Technologies and the Art of Trading-off," Proc. of the 2011 IEEE Infocom Work. On Green Comm. And Net. (GCN), Shangai, China, Apr. 2011.

[14] A multi-source dataset of urban life in the city of Milan and the Province of Trentino Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV.

[15] The 3GPP Association, "System Architecture for the 5G System," 3GPP Technical Specification (TS) 23.501, Stage 2, Release 16, version 16.6.0, Oct. 2020.

[16] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper.

[17] R. Bruschi, F. Davoli, C. Lombardo, O. R. Sanchez, "Evaluating the Impact of Micro-Data Center (μDC) Placement in an Urban Environment", 2018 IEEE Conf. on Network Function Virtualization and Software Defined Networks (IEEE NFV-SDN), Verona, Italy, 2018.

[18] https://opencellid.org/downloads.php.