Artificial Intelligence for Elastic Management and Orchestration of 5G Networks

David M. Gutierrez-Estevez, Marco Gramaglia, Antonio De Domenico, Ghina Dandachi, Sina Khatibi, Dimitris Tsolkas, Irina Balan, Andres Garcia-Saavedra, Uri Elzur, and Yue Wang

Abstract

The emergence of 5G enables a broad set of diversified and heterogeneous services with complex and potentially conflicting demands. For networks to be able to satisfy those needs, a flexible, adaptable, and programmable architecture based on network slicing is being proposed. A softwarization and cloudification of the communications networks is required, where network functions (NFs) are being transformed from programs running on dedicated hardware platforms to programs running over a shared pool of computational and communication resources. This architectural framework allows the introduction of resource elasticity as a key means to make an efficient use of the computational resources of 5G systems, but adds challenges related to resource sharing and efficiency. In this article, we propose Artificial Intelligence (AI) as a built-in architectural feature that allows the exploitation of the resource elasticity of a 5G network. Building on the work of the recently formed Experiential Network Intelligence (ENI) industry specification group of the European Telecommunications Standards Institute (ETSI) to embed an AI engine in the network, we describe a novel taxonomy for learning mechanisms that target exploiting the elasticity of the network as well as three different resource elastic use cases leveraging AI. This work describes the basis of a use case recently approved at ETSI ENI.

Introduction

In order to achieve the 5G Key Performance Indicators (KPIs), the most relevant standardization bodies have already defined the fundamental structure of the 5G architecture. By leveraging Software Defined Networking (SDN), Network Function Virtualization (NFV) and modularization, the new architecture proposed by relevant organizations such as the 3rd Generation Partnership Project (3GPP) or the European Telecommunications Standards Institute (ETSI) will natively support the service diversity targeted by the future commercial ecosystem [1, 2].

Besides the design of access and core functions, one of the most challenging tasks to be accomplished is network management. That is, the transition from the rather fixed operations support system/business support system (OSS / BSS) capabilities, to a new hierarchy of elements that have to deal with a very complex ecosystem of tenants, network slices, and services, each one with different requirements. In addition to management, 5G networks need orchestration capabilities that in turn are further divided into two main categories: service orchestration and resource orchestration. The former deals with the specific virtual network functions (VNFs) that compose a network slice, while the latter takes care of assigning resources to them. Tasks such as deciding whether a VNF shall be shared across slices or across tenants, their location in a possibly highly heterogeneous cloud infrastructure, or the number of allocated CPU cores are just a few examples of the Management and Orchestration (MANO) layer responsibilities.

The design of an efficient multi-service, multi-slice, and multi-tenant MANO entails challenges on both the architectural and algorithmic levels. Although the state-of-the-art MANO already provides baseline functionality, high computational resource efficiency is a real challenge today, and it is further aggravated by the complexity introduced by a 5G architecture based on the infrastructure sharing principle of network slicing. Our assertion is that an optimized utilization of cloud resources in the network, while providing the desired Service Level Agreement (SLA) under 5G network slicing, can only be achieved if fast and very fine-grained AI algorithms are designed and integrated into the network architecture itself. This allows for a more cost-efficient network management and orchestration by avoiding both resource under- and over-provisioning, which are the main causes of service outages and excessive expenditure, respectively.

Resource Elasticity

In order to solve the aforementioned problems, we have introduced the concept of resource elasticity for networks [3]. In a nutshell, the resource elasticity of a communications system can be defined as the ability to gracefully adapt to load and other system changes in an automatic manner such that at each point in time the available resources match the demand as closely and efficiently as possible. Furthermore, temporal and spatial traffic fluctuations in networks require effi-

David M. Gutierrez-Estevez and Yue Wang are with Samsung Research UK; Marco Gramaglia is with University Carlos III of Madrid; Antonio De Domenico and Ghina Dandachi are with CEA Leti France; Sina Khatibi is with Nomor Research Germany; Dimitris Tsolkas is with Mobics Greece; Irina Balan is with Noka Bell Labs Germany; Andres Garcia-Saavedra is with NEC Research Laboratories Europe GmbH Germany; Uri Elzur is with Intel HQ.

Digital Object Identifier:
10.1109/MWC.2019.1800498

cient network resource scaling: the network shall adapt its operation by eventually re-distributing the available resources as needed, up to the point of gracefully scaling the network performance to deal with excessive peak demand, thus avoiding abrupt decays. Although elasticity in networks has traditionally been exploited in the context of communications resources (e.g., when the network gracefully downgrades the quality for all users if communications resources such as spectrum are insufficient), here we address the computational aspects of resource elasticity since the virtualization and cloudification of networks at the core network (CN) and partially at the radio access network (RAN), means that the management and orchestration of its computational resources have now become a key challenge of 5G systems. In fact, in contrast with 4G systems, network slicing requires virtualized 5G networks to be able to jointly optimize communication and cloud resources.

We further consider elasticity in three different dimensions, namely *computational elasticity* in the design and scaling of VNFs; *orchestration-driven* elasticity achieved by flexible placement of VNFs; and *slice-aware elasticity* via cross-slice resource provisioning mechanisms. These dimensions encompass the full operation of the network and together they build our proposed elastic management and resource orchestration. To that aim, we envision a very prominent role for AI, as a tool to enhance the performance of elasticity algorithms. AI, and in particular machine learning (ML), has been proposed as a toolbox for different aspects of wireless networks [4]. In the context of elasticity, some examples of performance-boosting capabilities that could be provided by AI techniques are the following:

- Learning and profiling the computational utilization patterns of VNFs, thus relating performance and resource availability.
- Traffic prediction models for proactive resource allocation and relocation.
- Optimized VNF migration mechanisms for orchestration using multiple resource utilization data (CPU, RAM, storage, bandwidth).
- Optimized elastic resource provisioning to network slices based on data analytics.

Although by AI we refer to a wide range of techniques that could be employed for network management and orchestration, in this article we focus on three use cases that leverage specific ML algorithms, that is, drawing from a subset of the whole AI range of techniques, to exploit resource elasticity as follows:

- A computationally elastic scheduler applying deep learning to signal-to-noise ratio (SNR) prediction and the reinforcement learning technique of contextual bandits for making scheduling decisions.
- Slice-aware resource management based on traffic prediction using deep artificial neural networks (i.e., supervised learning).
- Efficient slice setup using the unsupervised learning technique of spectral clustering.

It is worth mentioning that even though these three specific examples of AI-based elasticity algorithms utilize ML techniques, the authors believe that other AI techniques, not necessarily constrained to the ML domain, could also be applied.

The remainder of this article is structured as follows. In the following section, we provide a description of a prominent architecture for the use of AI in the management and orchestration of future networks proposed by ETSI. Then we discuss the application of AI in the context of resource elasticity and we elaborate on the above mentioned elasticity use cases and the AI techniques they employ. Finally, we conclude the article.

AI-Enabled 5G Network Architecture

In response to the industry demand for AI-driven intelligent networks, ETSI has created the ENI work group [5]. ENI's goal is to improve the operator's experience and add value to telco provided services, by assisting in decision making to deliver operational expenditure (OPEX) reduction and to enable 5G deployment with automation and intelligence. In particular, ENI aims to define an architecture that uses AI techniques and context-aware, metadata-driven policies, to adjust service configuration and control based on changes in user needs, environmental conditions, and business goals, according to the "observe-orient-decide-act" control loop model [5].

Network slicing for 5G can serve as a prime example to demonstrate ENI's architecture and the operator's benefits it provides, especially around VNF's computational resource efficiency, while preserving the user requested SLA.

The telco industry's evolution toward standardization of AI-assisted networks requires various industry consensus, including grammar and syntax for service policy and associated domain specific language (DSL), as well as data ingestion format, to foster the ability to interact with the broad variety of tools used for management and monitoring. A *normalized* format is required also to address the difficulty to harmonize the state of the divergent infrastructure, due to the use of silo specific tools, for example, per compute, network, and storage, and due to the variety of "assisted systems," each with different capabilities and different exposed API and varying degrees of ability to interact with the AI system, like ENI. It is therefore essential for ENI to define architecture components such as data ingestion and normalization, to provide a common base for ENI's inter-modular interaction as well as for transforming the external assisted system (e.g., a 3GPP/5G implementation) inputs to a format that is understood by ENI.

To date, ENI has defined a modularized system architecture, as shown in Fig. 1a. Having a modularized system architecture facilitates the flexibility and generalization in the system design, and increases vendor neutrality. A brief description of each module, according to [5], is given below.

The Policy Management Module: Provides decisions to ensure that the operator's goals and regulator's policies are met.

The Context Awareness Module: Describes the state and environment in which a set of the assisted system entities exists or has existed. For example, an operator may have a business rule that prevents 5G from a specific type of network slice in a given location.

The Situational Awareness Module: Enables ENI to understand how information, events, and

The telco industry's evolution toward standardization of AI-assisted networks requires various industry consensus, including grammar and syntax for service policy and associated domain specific language (DSL), as well as data ingestion format, to foster the ability to interact with the broad variety of tools used for management and monitoring.

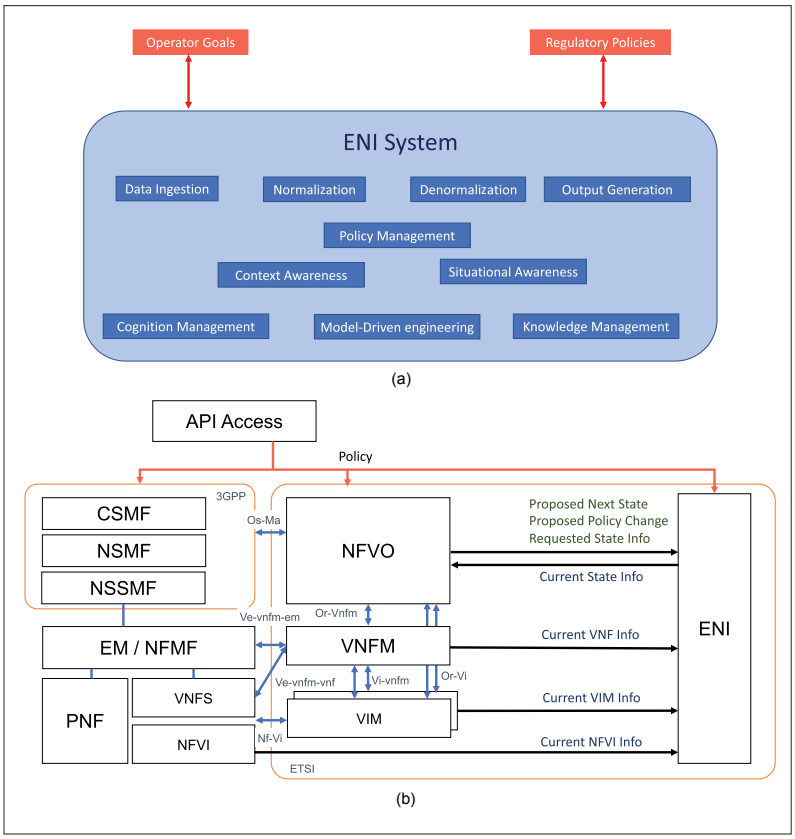

FIGURE 1. The ETSI ENI Architecture and its interaction with the ETSI NFV MANO framework: a) ETSI ENI modularized system architecture; b) Management and orchestration Architecture.

recommended commands that it may provide to the assisted system may impact its next state, actions, and ability to meet its operational goals.

The Cognition Management Module: Operates at the higher level and enables ENI as a whole to consult and meet its end to end goals.

The Knowledge Management: Used to represent information about ENI and the assisted system, differentiating between known facts, axioms, and inferences.

The interaction and interoperability of ENI with an assisted system is determined by the latter's support of the ENI Reference Points [5]. Specifically for the use of compute resources elasticity and efficiency, as presented in this article, few elements, determined by relevant ENI Reference Points are needed. As depicted in Fig. 1b, the current NFVI Information allows ENI to be aware of the computational resources' capabilities (e.g., type of CPU, memory, data plane and accelerators) and availability (status and utilization level), while in turn this enables ENI to influence and optimize placement decisions made by the VIM, while ensuring that 3GPP policies, resource allocation and SLA are adhered too. Moreover, by using this information, ENI can further optimize resource utilization by:

- Enabling higher density for a given set of workloads under associated SLA.
- Anticipating and reacting to changing loads in different slices and assisting the VIM in avoiding resource conflicts.
- Timely triggering of up/down scaling or in/out scaling of associated resources.

Applying AI in Softwarized Mobile Networks: A Taxonomy View

Despite recent publications in the field [6], the full integration of AI in mobile network architecture is still in its early stages, and the design of learning algorithms that provide promising features such as network elasticity, as described earlier, is still a greenfield research topic. In this section, we describe learning techniques for applying and exploiting elasticity in the upcoming generations of mobile networks. Specifically, we propose a taxonomy on the learning characteristics required to provide elasticity, and three specific AI-based elasticity use cases, namely elastic RAN VNF design, slice-aware elastic resource management, and efficient slice setup.

We propose two different taxonomies for learning in the context of elasticity based on the data used for learning, and the network slice lifecycle phase. First, with respect to the data, learning techniques for elastic network slice management can be categorized along two main directions, as described below, independently of the actual algorithm in place:

Inputs: Learning techniques shall learn features from the user demand to the network, the infrastructure utilization and the slice policies. These inputs shall be conveniently measurable (and labeled in case of supervised techniques) in order to be applied in one of the outputs.

Outputs: Following the 3GPP definition [7], lifecycle management is composed of four stages: preparation, instantiation, run-time and decommissioning. Hence, depending on the kind of algorithms, its target and the input features, the learning algorithm shall be employed in one of these phases.

The input direction can be further split along three dimensions, depending on the characteristics of the learned input feature. In Fig. 2 we show this three-dimensional classification, highlighting its three main axes: the *demand*, the *infrastructure*, and the *requirements*. Triangles in Fig. 2 represent the granularity on each of the axes, being the darker the finer.

Demand: Learning user behavior is paramount for enforcing elasticity in the network. As previously discussed, the multiplexing gains achieved by efficiently combining different slices on the same infrastructure necessarily requires learning of the user demand. That is, anticipatory resource re-orchestration builds on the understanding of the temporal and spatial demands of services. This input data may have a coarse granularity (i.e., order of minutes) as the current orchestration technologies and the increased signaling overhead caused by numerous re-configurations prevent a too fast resource reassignment. This operational point is marked as D2 in Fig. 2. Nevertheless, demand may be learned at more granular levels (D1 in Fig. 2) when designing elastic RAN NFs. In this case, learning metrics such as the user requests queuing reports at faster time scale (i.e., sub-seconds) enables better decision making on the short-term future scheduling decisions according to the available computation capacity.

Infrastructure: Learning how the underlying infrastructure reacts or limits elastic management/orchestration decisions is fundamental. For exam-

ple, elastic resource assignment algorithms need to learn about the computational behavior of NFs when subject to a certain load and to different requirements to provide a precise VNF location (I1 in Fig. 2). Analogously, the wireless infrastructure (i.e., the channel) is probably the main driver for the elastic behavior of RAN functions, as it is the most important limiting factor.

Requirements: A very important challenge for future sliced 5G networks is the service creation time. ML can greatly enhance the service setup by automatically translating consumer-facing service descriptions into resource-facing service descriptions that can be processed by the network management and orchestration functionality in order to allocate the proper resources to the new service. AI tools can thus replace human interventions, which increase costs and are time consuming, to identify the resource requirements of a new service from the slice down to the VM/container levels. Furthermore, this approach can smartly take into account existing services with similar requirements to favor resource multiplexing across services and increase the system efficiency.

On the output dimension, the proposed taxonomy refers to the network slice lifecycle phases, as various approaches can be adopted and applied in all the phases of the lifecycle of a slice instance [7]. For example, slice behavior analysis can be a critical asset for elasticity provisioning in the slice preparation phase, since statistics can be exploited to efficiently decide the basic configurations and set the network environment.

In this article, we provide insights and use cases on AI-based elasticity mechanisms that are applied in the instantiation and run-time phases, but the preparation and decommissioning phases could similarly benefit from AI.

Instantiation Phase: The pool of parameters that feed the learning process of AI-based elastic mechanisms in this phase may be:

- Requirements depicted in SLAs and service demands.
- Past measurement and statistics related to resource consumption profiles of VNFs.
- Real time measurements from already instantiated slices.
- The current state of computational and resource consumption in the system.

Based on these factors, the AI mechanism decides the admission of new slices and potentially the re-configuration of the running slices in the network. Here, we focus on slice setup mechanisms based on AI that guarantee flexible slice admission control and deep network slice blueprint or template analysis. Later we propose a learning approach for network slice admission control, which precisely takes place in the instantiation phase.

Run-Time Phase: For the AI-based elasticity mechanisms that are applied in the slice run-time phase, all the parameters that are available in the instantiation phase can be exploited. However, the learning capability is much more challenging since traffic load measurements are available, while the adaptation should be done in a faster scale, including re-configurations at the VNF or slice level. Here, we focus on advanced sharing of computation resources among VNFs of multiple slices to provide resource elasticity, while the involved slices are in operation. Such an approach is presented below.

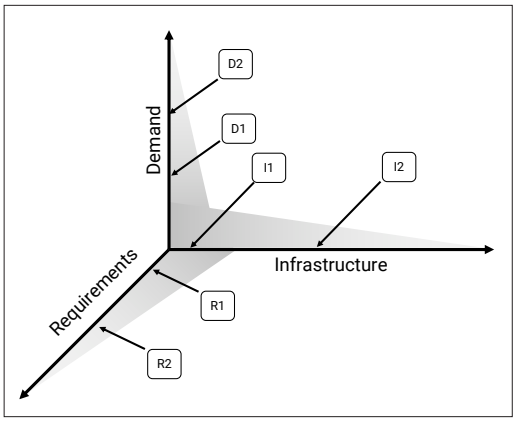

FIGURE 2. Learning taxonomy axes for slice lifecycle management.

Furthermore, the challenge of enabling VNF self-adaptation during the run-time phase is handled.

Challenges

The above taxonomy is useful to understand where AI can help in the management and orchestration of networks. However, the selection of the right AI-based algorithm is not necessarily a trivial task. Clearly, the features of the learned parameters described in this taxonomy do have an impact on the type of learning algorithm that is employed. For example, highly dynamic parameters such as load may require algorithms with fast and adaptive online learning capabilities; yet other parameters such as the slice blueprint given the service requirements are more static and offline training could suffice for an artificially intelligent system to make the right decisions. Hence, although the fast-evolving field of AI makes difficult an a-priori selection of certain types of learning algorithms (e.g., deep neural networks, reinforcement learning, and so on) for specific types of parameters, it becomes apparent that a correlation between those does exist, and the design of the learning system and algorithms must carefully take into consideration such a correlation. In addition, labeled (and reliable) data sets to implement supervised learning algorithms in many cases are only (partially) available, since 5G deployment is not started yet. Furthermore, these AI algorithms may deliver but a sliver of the more comprehensive and ambitious goals of cognitive network management systems where architectural support is also required. An analysis of such architecture requirements is, however, out of the scope of this article, but the interested reader is referred to [8], where extensive architectural impact analysis has been performed.

Use Cases

Next, we describe three possible use cases for the application of AI algorithms that target network elasticity by applying cognitive techniques on different inputs and in different phases of the lifecycle.

Computationally Elastic Scheduler

As discussed earlier, computational elasticity deals with the performance optimization of a NF given additional constraints on the available computational capacity assigned to such a function by an

Learning how the underlying infrastructure reacts or limits elastic management/orchestration decisions is fundamental. For example, elastic resource assignment algorithms need to learn about the computational behavior of NFs when subject to a certain load and to different requirements to provide a precise VNF location.

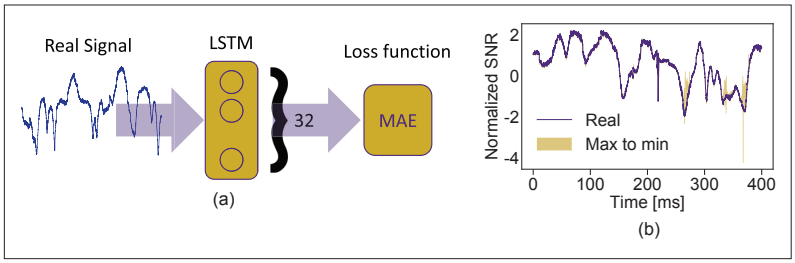

FIGURE 3. A framework for SNR prediction: a) NN architecture; b) SNR prediction.

orchestrator. While this approach can be applied to any kind of NF, those that imply higher computational loads can benefit more. As consistently shown in the literature, the most expensive NFs in terms of computational demand are the ones related to the MAC, encoding and decoding [9]. In a previous work [10], we proposed an algorithm for uplink MAC scheduling that offers graceful degradation in case of a sudden load variation that could not be served with the available computational resources (i.e., a flash crowd).

The algorithm in [10] showed indeed the potential of such an approach. However, it relies on strong assumptions on both the channel conditions and the user demand. As already discussed, such metrics shall be considered as part of a learning process. In the following, we propose a possible approach to an AI-based computational elasticity.

In nuce, achieving computational elasticity at the MAC level implies a joint optimization of Modulation and Coding Scheme (MCS) selection for each user (as discussed in [11], different MCS have different computational complexity depending on the SNR margin), and the actual user scheduling. That is, rather than only selecting who to schedule, the elastic MAC controller shall also select the best MCS to be used given the constraints on the available computational capacity.

Thus, selecting the best scheduling decision at each time transmission interval (TTI) entails learning characteristics such as the traffic demand and the channel conditions. However, given the trend of centralizing access NFs, it is likely that an elastic MAC scheduler will need to take scheduling decision for thousands of devices at the same time. Therefore, the scalability of the learning process is of paramount importance for its practical implementation. A promising learning solution for solving this problem is that of contextual bandits [12]. Contextual bandits employ the concept of policy selection, as opposed to action selection in classical bandit problems. A policy essentially maps context information (encoded as a sample from a potentially rich feature space) into a scheduling action. By learning the history of policy-context-reward tuples, randomized greedy algorithms can be built to maximize the total reward for any upcoming context, which in this case includes the user data queues and the buffer state of the computing processor.

A necessary input for contextual bandits is, as discussed previously, the prediction of the infrastructure status for a given time frame. In mobile networks, forecasting the SNR quality of a given user is, thus, fundamental to take the scheduling decisions as described above. We thus explored the feasibility of a SNR prediction algorithm (results are depicted in Fig. 3). The objective was to obtain a short scale (5 ms) forecast of the SNR values, taking into account a window of the past 40 ms samples. For this purpose we employ a layer of a Long Short Term Memory (LSTM) network, activated with a Scaled Exponential Linear Unit (SELU) function and a Mean Absolute Error (MAE) loss function (Fig. 3a). As shown in Fig. 3b, this network is capable of forecasting a real world SNR trace collected in a lab environment, demonstrating the effectiveness of a learning scheduling framework.

Slice-Aware Resource Management

The design and setup of a network slice capable of accurately satisfying the need of mobile services with very diverse requirements is an important challenge for 5G networks. This process can be optimized by enabling the 3GPP Network Slice Management Function (NSMF) and Network Slice Subnet Management Functions (NSSMF) to use AI mechanisms capable of automatically translating service requirements to network requirements. To this aim, 3GPP recently introduced the Management Data Analytics Service (MDAF) in the orchestration architecture [13].

The goal in slice-aware elastic resource management is to develop algorithms, which consider the Quality of Service (QoS) requirements, SLAs, and demands of network slices operating on the same physical infrastructure to optimally allocate/de-allocate a portion of available resources to each of them. The two main design challenges are modeling of the essential parameters, and adapting the models to changes in the run-time. This information is extremely useful for resource allocation and provisioning at every level of the network. In a scenario where a limited number of RAN radio resources have to be shared among multiple slices with significantly different requirements, different RAN parameter set configurations are needed. These may vary in spatial domain due to changing radio conditions as users are moving, and in temporal domain depending on the traffic load distribution over time.

The *VNFs computational performance* is highly dependent on the implementation techniques as well as channel quality. In [14], a profiling procedure has been proposed; it uses AI-based regression (i.e., Lasso regression), to generate a mathematical model. On the same research path, AI-based solutions (Lasso, Support Vector Machine (SVM), or reinforcement learning) can learn (or adapt) the computational performance of VNFs based on the reported input parameters and the measured processing times for any new VNFs.

The *channel quality* between the antennas and the mobile terminals is the foundation to estimate the total network throughput and allocate the available radio resources to each slice as well as the required computational resources. In both cases, AI-based approaches can either provide or adapt the channel models based on the monitoring reports to be used in estimations and provisioning of slice-aware resource management algorithms.

AI techniques could also be used for *traffic prediction*, which can be a valuable input for many elastic resource allocation algorithms. The resource management algorithms either act in passive mode

(i.e., observing the demand and react to it) or always assume the maximum demand. The prediction of slice demands can enhance inter-slice resource utilization. Figure 4a presents the deep neural network architecture with two dense layers with ReLU activation function and a sigmoid activation function. It is used to predict the traffic demands of two network slices with different behaviors, and Fig. 4b shows the predicted against the actual traffic. Virtual resource management models, consequently, can now also consider predicted slice demands to adapt the service provisioning; for example, some services may have a repetitive pattern or may only be active during certain times of the day or year.

The movement of traffic concentration around the network could also be predicted, for example, groups of users could be identified that move in a coordinated fashion through the network and following a certain trajectory. Such input may be very useful for *adjusting the beam patterns* of groups of cells proactively. Dynamic beam pattern adjustment would shift the load distribution between cells and ensure that all users are best served at the same time. Knowing in advance the traffic characteristics of each slice and its evolution over time and space is essential to reaching the correct beam forming for each cell and aligning across neighbors in order to create stable coverage in a timely manner. This is clearly valuable for latency sensitive services/slices or throughput-hungry ones.

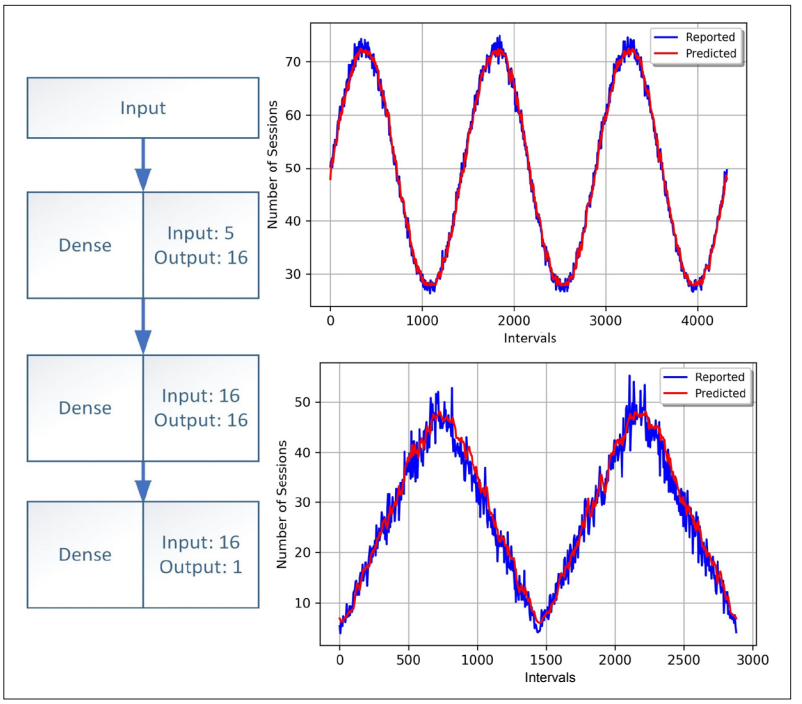

FIGURE 4. Traffic demand prediction using deep neural network.

Efficient Slice Setup

We envision an important role of AI algorithms during the run-time phase of a network slice. However, unsupervised learning algorithms are fundamental also in the instantiation phase, where they shall analyze the generated requirements and identify whether a slice already instantiated can efficiently support the new service or an additional slice needs to be deployed. This approach not only further reduces the service creation time by avoiding the instantiation of a new slice for each new service, but also enhances the system efficiency by increasing the resources shared across elastic slices. To be effective, this approach has to operate on slices that do not need fully dedicated resources, for example, they are elastic in the sense that they have relaxed constraints in terms of resource isolation. In contrast, slices characterized by stringent resource isolation constraints are non-elastic and may not accept to share their resources with concurrent slices and limit the system flexibility.

A practical example is the case of different broadcasters covering the same sporting event: 3GPP's NSMF may mutualize the radio resources allocated to the different services to transmit common content, and use dedicated resources for slice-specific content such as the speaker's voice. More specifically, most mobile services are typically characterized by a set of dedicated NFs in charge of guaranteeing its specific requirements (e.g., multi-connectivity for high reliability) and a larger set of shared NFs that deal with more generic requirements (e.g., the handover function that guarantees coverage).

An AI-based mechanism can classify in an unsupervised manner the instantiated slices with respect to the NFs shared with the new request, and then assign the new slice request to the deployed slice based on the number of shared NFs. In this way, the additional resources needed to fulfill the requirements of the new slice can be reduced and the slice deployment process accelerated. This approach can also be used as a congestion control mechanism to prevent resource outages: when the system is close to saturation, the NSMF can re-cluster the overall set of services in new network slice instances to maximize the resource sharing. The latter could be implemented by using a *spectral clustering scheme* [15], where the deployed slices are represented as nodes of a connected graph and clusters are found by partitioning this graph based on the nodes' affinity (e.g., related to the number of shared NFs). Figure 5 shows the variation of the slice request dropping probability as a function of the non-elastic slice arrival probability. In this result, slices are classified between elastic and non-elastic and we assume that non-elastic slices lead to high revenues as they require dedicated network resources. We evaluate the performance of three different approaches. In the first approach, resource sharing is not implemented, which results in higher resource requirements and larger slice dropping probability. In the second case, we assume that resource sharing is enabled by assigning a new slice request to the already instantiated slice maximizing the number of common VNFs (i.e., max VNF). Finally, in the third case, spectral clustering is implemented at each slice request arrival/departure to maximize the resource sharing in the system. Spectral clustering shows the best performance since it continuously optimizes the shared resources at the cost of higher complexity. The results in Fig. 5 show that both mechanisms enabling resource sharing improve the performance for both elastic and non-elastic slices; however, the slice dropping probability reduction obtained when using the simple max VNF approach is limited (around 11 percent). In

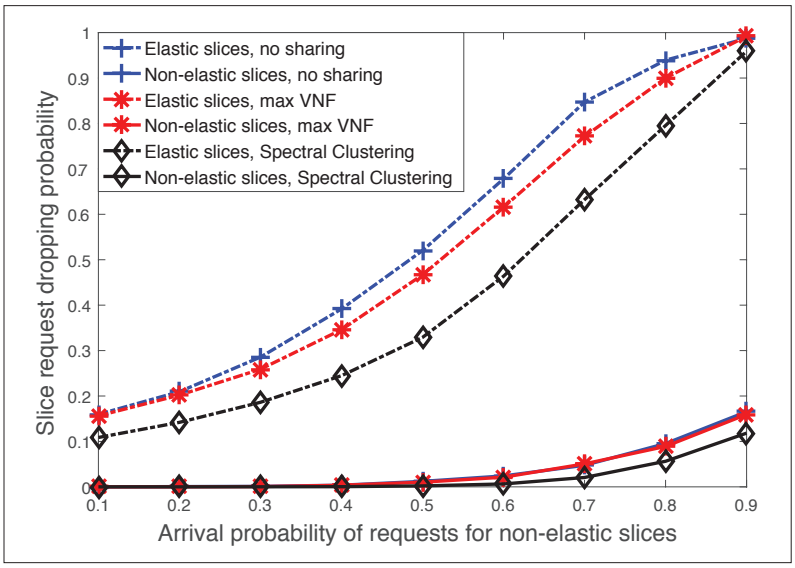

FIGURE 5. Elastic and non-elastic slice dropping probability as a function of non-elastic slice arriving probability with and without resource sharing.

contrast, the spectral clustering approach leads to 50 percent reduction of the slice dropping probability, therefore enabling a large improvement in terms of potential income for the operator.

Conclusions

In this article, we have introduced the novel idea of utilizing AI techniques with the purpose of exploiting the resource elasticity of a 5G network, hence improving resource efficiency and the overall performance of its management and orchestration machinery. Using as a basis the architectural work recently developed by ETSI ENI and the concept of resource elasticity, we propose a taxonomy for elastic slice lifecycle management and three different use cases showing the applicability of AI on different management and orchestration problems where elasticity can be exploited. The article constitutes the basis of a recently approved use case at ETSI ENI.

Acknowledgment

Part of this work has been performed within the 5G-MoNArch project (Grant Agreement No. 761445), part of the Phase II of the 5th Generation Public Private Partnership (5G-PPP) program partially funded by the European Commission within the Horizon 2020 Framework Program. This work was also supported by the the 5G-Transformer project (Grant Agreement No. 761536).

References

[1] 3GPP, "System Architecture for the 5G System," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501, 09 2018, version 15.3.0.
[2] ETSI, "GS NFV-IFA 014 — Network Functions Virtualisation (NFV); Management and Orchestration; Network Service Templates Specification," Tech. Rep., Oct. 2016.
[3] D. M. Gutierrez-Estevez *et al.*, "The Path Towards Resource Elasticity for 5G Network Architecture," *Proc. 2018 IEEE Wireless Commun. and Networking Conf. Workshops (WCNCW)*, Barcelona: IEEE, Apr. 2018, pp. 214–19.
[4] M. Chen *et al.*, "Machine Learning for Wireless Networks with Artificial Intelligence: A Tutorial on Neural Networks," arXiv preprint arXiv:1710.02913, 2017.
[5] Y. Wang *et al.*, "Network Management and Orchestration Using Artificial Intelligence: Overview of ETSI ENI," *IEEE Commun. Standards Mag.*, vol. 2, no. 4, 2018, pp. 58–65.
[6] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Commun. Surveys Tutorials*, 2019.
[7] 3GPP, "Telecommunication Management; Study on Management and Orchestration of Network Slicing for Next Generation Network," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 28.801, 01 2018, version 15.1.0.
[8] 5G Mobile Network Architecture for Diverse Services, Use Cases, and Applications in 5G and Beyond, 2017; available: https://5g-monarch.eu/
[9] A. Garcia-Saavedra *et al.*, "WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul," *IEEE Trans. Mobile Computing*, vol. 17, no. 10, Oct. 2018, pp. 2452–66.
[10] P. Serrano *et al.*, "The Path Toward a Cloud-Aware Mobile Network Protocol Stack," *Trans. Emerging Telecommunications Technologies*, vol. 29, no. 5, May 2018, p. e3312.
[11] P. Rost, S. Talarico, and M. C. Valenti, "The Complexity–Rate Tradeoff of Centralized Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, Nov. 2015, pp. 6164–76.
[12] A. Agarwal *et al.*, "Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits," *Proc. Int'l. Conf. Machine Learning*, Jan. 2014, pp. 1638–46.
[13] 3GPP, "Management and Orchestration; Architecture Framework," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 28.533, 09 2018, version 15.0.0.
[14] S. Khatibi, K. Shah, and M. Roshdi, "Modelling of Computational Resources for 5G RAN," *Proc. 2018 European Conf. Networks and Communications (EuCNC)*, Ljubljana, Slovenia: IEEE, June 2018, pp. 1–5.
[15] S. E. Schaeffer, "Graph Clustering," *Computer Science Review*, vol. 1, no. 1, Aug. 2007, pp. 27–64.

Biographies

David M. Gutierrez-Estevez is a principal research/standards engineer at Samsung Electronics R&D Institute UK. He obtained his Engineering Degree in telecommunications (Hons.) from the Universidad de Granada, Spain, and his M.S. and Ph.D. degrees from Georgia Institute of Technology in Atlanta, GA, USA, the latter supported by graduate fellowships from Fundacion la Caixa and Fundacion Caja Madrid in Spain. He obtained the Broadband Wireless Networking Lab Researcher of the Year Award in 2013 for outstanding research contributions during his Ph.D. From September 2014 to September 2015 he worked for Huawei Technologies in Silicon Valley. Prior to that, he held an internship position at the Corporate R&D Division of Qualcomm in San Diego, CA, USA, as well as research assistant and intern positions at Fraunhofer Heinrich Hertz Institute and Fraunhofer Institute for Integrated Circuits, both in Germany. In January 2016 he joined Samsung UK, where he has led numerous research activities in several 5G-PPP research projects over the span of three and half years. Since 2019, he has been a 3GPP delegate within SA2 (system architecture). His published work has accumulated over 1000 citations, his H-index is 12, and he is a co-inventor of several patents and patent applications.

Marco Gramaglia is a post-doc researcher at the University Carlos III of Madrid (UC3M), where he received M.Sc (2009) and Ph.D. (2012) degrees in telematics engineering. He held post-doctoral research positions at the Istituto Superiore Mario Boella (Italy), the Institute of Electronics, Computer, and Telecommunications Engineering (IEIIT) of the National Research Council of Italy (Italy), and IMDEA Networks (Spain). He was involved in EU projects and has authored more than 40 papers published in international conferences and journals.

Antonio De Domenico received his M.Sc. and Ph.D. degrees in telecommunication engineering in 2008 and 2012 from the University of Rome "La Sapienza" and the University of Grenoble, respectively. Since 2009, he has worked with the CEALETI – MINATEC, Grenoble, France, as a research engineer. His research topics are cloud enabled heterogeneous wireless networks, millimeter-wave based communications, machine learning, and green communications. He is the main inventor or coinventor of 12 patents. In 2017, he was awarded by the CEA Enhanced Eurotalents Programme. In 2018, he was a visiting researcher in the Communications Group of the Department of Electrical and Computer Engineering at the University of Toronto.

Ghina Dandachi received the M.S. degree in telecommunication system technologies and the M.E. degree in computer science and telecommunications in 2013 from the Lebanese University, Beirut, Lebanon. She received the Ph.D. degree jointly at Telecom SudParis, Paris, France, the University of Paris VI,

Paris, and the Lebanese University in 2017. Since 2017, she has worked as a research engineer at CEA LETI. Her research interests include artificial intelligence mechanisms for dimensioning and performance optimization in wireless networks.

Sina Khatibi is with NOMOR Research GmbH, Munich, Germany as a project leader and senior researcher, leading Nomor's research team in EU research projects. His main research focus is Machine Learning (ML) and deep learning approaches for network slicing and Self-Organized Networks (SONs) optimization in addition to analytical modelling of networks for long-term planning. He received his Ph.D. from the Electrical and Computer Engineering program of IST, the University of Lisbon, Portugal, in 2016.

Dimitris Tsolkas holds a Ph.D. degree from the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens (NKUA). He has extensive experience in R&D and project management, working in a plethora of EC-funded research and development projects in association with institutions from academia (Department of Informatics and Telecommunications - NKUA, Computer Technology Institute and Press "Diophantus", and Department of digital systems - University of Piraeus) and industry (LINK Technologies S.A., WEST L'Aquila, and MOBICS S.A.). He has teaching experience as a lecturer at the Business College of Athens (BCA) and as an instructor of postgraduate and graduate courses in the Department of Computer Science and Engineering, University of Ioannina, and the Department of Informatics and Telecommunications – NKUA. He has published more than 35 articles in peer-reviewed journals, international conferences and book chapters. His research interests are in the areas of 5G RAN, D2D communications, and QoE provisioning.

Irina Balan received her Master's degree in telecommunication engineering from "Gheorghe Asachi" Technical University, Iasi, Romania in 2007, and her Ph.D. in computer science from Ghent University, Belgium in 2012. Since 2013, she has been with Nokia Bell Labs Research in Munich, Germany. Her current research interests include 5G/New Radio topics such as beam management, SON (Self Organizing Networks) from mobility mechanisms and oredictive traffic patterns via Machine Learning.

Andres Garcia-Saavedra received his M.Sc. and Ph.D. from University Carlos III of Madrid (UC3M) in 2010 and 2013, respectively. He then joined the Hamilton Institute, Ireland, as a research fellow until the end of 2014, when he moved to Trinity College Dublin (TCD). Since July 2015 he has been a senior researcher at NEC Laboratories Europe. His research interests lie in the application of fundamental mathematics to real-life communications systems and the design and prototype of wireless systems and protocols.

Uri Elzur is the CTO for the Data Center Network Solution Group (DNSG) of Intel's Data Center Group. In this role, he is responsible for creating SDN/NFV technical strategy, open source top-to-bottom stack architectures (including MANO, OpenStack, ODL, vSwitch) and influencing the server platforms. Uri is a networking specialist with more than 25 years of industry experience and a proven track record of creating innovative product architectures, strategies and intellectual property in networking, security and related technologies. Prior to joining Intel, Uri held a position of a senior director at Broadcom, managing an architecture team with responsibilities over the company's NIC architecture and strategy. In that role Uri led multiple innovations in the areas of virtualization, TCP offload, RDMA, iSER, iSCSI/FCoE and more. Uri holds many patents, represented his employer in multiple standards organizations and industry consortiums including Open-O (as VC Architecture), OpenDayLight, ONF, OpenStack, IETF, IEEE, T.11, DMTF and RDMA Consortium, and is the co-author of a few RFCs. His experience in public speaking includes multiple presentations at OpenStack, the ODL summit, VMworld and more recently the Intel Developer Forum, MEF and Layer123 SDN events. Uri holds a BSc EE and MSc EE/CS degrees from the Technion in Haifa, Israel.

Yue Wang is a principal 5G researcher at Samsung Electronics R&D Institute UK. Her current research focuses on AI for 5G, with topics spanning extensively on the application of AI in communications systems and networks for 5G and beyond. She is the Samsung delegate to ETSI ISG ENI (Experiential Networked Intelligence), and the Secretary and Rapporteur of ENI. She also sits on the Industry Advisory Board of two universities, and is the industry supervisor of a five-year research program, all in the area of AI for 5G and beyond. Prior to joining Samsung, she worked in the US and UK on a number of technical subjects in wireless communications research and standards. She received her Ph.D. from the University of Victoria. She has co-authored over 40 papers and is a (co)-inventor of over 20 patents. She has been a Senior Member of IEEE since 2012.