Network Intelligence in 6G: challenges and opportunities

Albert Banchs
University Carlos III of Madrid
IMDEA Networks Institute
Leganes, Spain
banchs@it.uc3m.es

Marco Fiore
IMDEA Networks Institute
Leganes, Spain
marco.fiore@imdea.org

Andres Garcia-Saavedra
NEC Laboratories Europe
Heidelberg, Germany
andres.garcia.saavedra@neclab.eu

Marco Gramaglia
University Carlos III of Madrid
Leganes, Spain
mgramagl@it.uc3m.es

ABSTRACT

The success of the upcoming 6G systems will largely depend on the quality of the Network Intelligence (NI) that will fully automate network management. Artificial Intelligence (AI) models are commonly regarded as the cornerstone for NI design, as they have proven extremely successful at solving hard problems that require inferring complex relationships from entangled, massive (network traffic) data. However, the common approach of plugging 'vanilla' AI models into controllers and orchestrators does not fulfil the potential of the technology. Instead, AI models should be tailored to the specific network level and respond to the specific needs of network functions, eventually coordinated by an end-to-end NI-native architecture for 6G. In this paper, we discuss these challenges and provide results for a candidate NI-driven functionality that is properly integrated into the proposed architecture: network capacity forecasting.

CCS CONCEPTS

• **Networks → Network design principles**.

KEYWORDS

Network Intelligence, Mobile Networks, 6G

ACM Reference Format:
Albert Banchs, Marco Fiore, Andres Garcia-Saavedra, and Marco Gramaglia. 2022. Network Intelligence in 6G: challenges and opportunities. In *16th ACM Workshop on Mobility in the Evolving Internet Architecture (MobiArch'21), January 31-February 4, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477091.3482761

1 INTRODUCTION

The vision for the sixth-generation (6G) mobile systems [7] sets an extraordinarily high bar for mobile networks, which are expected to become general-purpose platforms and provide Smart Connectivity to a plethora of extremely heterogeneous terminals. 6G shall support entirely diverse classes of services, and do so with outstanding performance: near-zero latency, apparent infinite capacity, and 100% reliability and availability will make the communication infrastructure fully transparent to the applications. Meeting this ambitious goal requires growing the already substantial complexity of mobile network architectures to instantly orchestrate physical resources and Virtual Network Functions (VNFs) across different network domains, in concertation with time-varying user demand and multi-tenancy requirements.

Managing the increased complexity of 6G networks with traditional human-in-the-loop approaches will not be possible anymore. Instead, technologies that fully automate the network operation will become the standard, and the success of 6G will vastly depend on the quality of the *Network Intelligence* (NI) that will run at schedulers, controllers, and orchestrators across network domains, and *de-facto* manage the infrastructure. More specifically, NI comprises the whole set of smart algorithms that will be deployed in 6G networks for automated decision making, adapting network resources and functions to the time-varying demand and performance requirements without a need for human intervention.

As in many other research and engineering domains, Artificial Intelligence (AI) implemented with complex machine learning models is regarded as a promising approach to design NI solutions. AI models have proven remarkably effective at addressing complex tasks, and they thrive on the large amount of control and traffic data available within network architectures. Indeed, there are multiple examples of NI tasks in which AI techniques excel, by finding complex relationships in vast traffic data and using them to support effective network operation (see [13] for an extensive review).

While early efforts such as those above are compelling, they are obliged to fit into current network management models, which were not designed to accommodate AI solutions. Instead, we argue that a sound integration of AI into the networking landscape will necessarily entail reshaping the network operation for NI support, as detailed in Section 2.

Specifically, in this paper we identify and explore two key challenges that need to be tackled towards structured incorporation of NI in 6G systems:

MobiArch'21, January 31-February 4, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-8706-4/22/01…$15.00
https://doi.org/10.1145/3477091.3482761

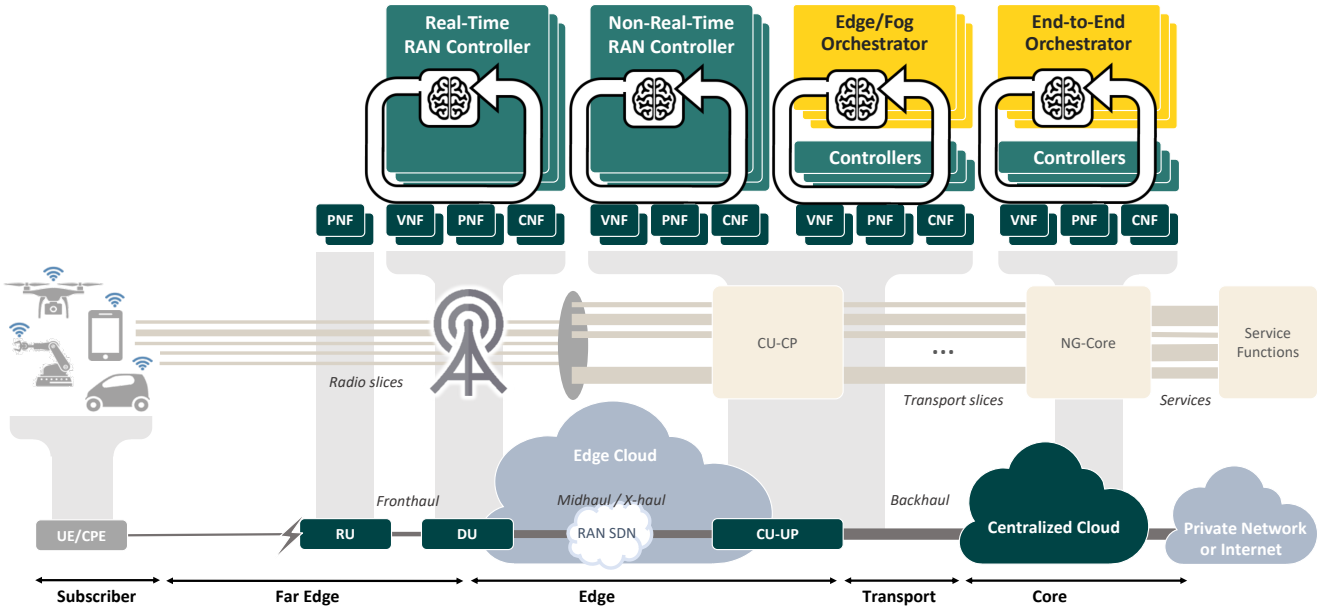

Figure 1: High-level view of the 6G network architecture envisioned by current standardization at O-RAN [11], 3GPP, and ETSI. "Brains" denote the NI instances located in network orchestrators and controllers, which interact with NFs via closed loops.

- **Design of an end-to-end NI-native architecture.** Practical NI algorithms shall be supported by a network architecture that is explicitly intended to host and facilitate the integration of NI. Such *NI-native* architecture shall stem from current standardization trends, and enable the coordination of the many and varied NI instances deployed across network domains. In particular, this architecture shall: (*i*) go beyond centralized orchestration, acknowledging that different network functionalities have major reciprocal effects and NI algorithms do not operate in a vacuum; and, (*ii*) provide NI directly at Network Function (NF) level. These points are discussed in detail in Section 3.
- **Development of customized AI techniques that empower practical NI.** In order to operate in the most effective way, AI models shall be tailored so as to respond to the specific needs of network management functionalities. To this end, AI models must take advantage of the most recent advances in machine learning to address overlooked design aspects that are fundamental for NI. Dedicated loss functions, latency guarantees, or reduced requirements in terms of training and computational complexity are all relevant aspects, as analyzed in Section 4.

The two challenges above consider complementary approaches. On the one hand, a NI-native architecture follows a strategy of updating the existing network design to best accommodate AI. On the other hand, the customization of AI looks at how to best adapt machine learning models to the needs of network functionalities. The convergence of the two approaches ultimately leads to efficient development, implementation, and integration of NI in 6G networks, so as to meet their very high expectations in terms of automation, performance, sustainability, and reliability.

2 NETWORK MANAGEMENT TRENDS AND CHALLENGES

In the transition towards their sixth generation, mobile networks will undergo an architectural revolution, aimed at supporting the extreme requirements set by future services that will assume performance indicators like virtually infinite capacity or perceived zero latency. Current efforts in standardization reflect this trend and are pushing a number of distinctive novelties into the mobile network design, which are summarized in Figure 1 and detailed below.

2.1 Mobile Networks architectural trends

The mobile network architecture is being redesigned for end-to-end softwarization and cloudification, completing the decoupling of NFs from the underlying hardware, and granting unprecedented capacities to control system-wide operations. As shown in Figure 1, a full range of VNFs and containerized Cloud-native Network Functions (CNFs) will complement traditional Physical Network Functions (PNFs). In this scenario, a variety of overarching entities are responsible for the orchestration (*i.e.*, service and network functions lifecycles management, including their instantiation, scaling, or termination) and control (*i.e.*, the parameter tuning of deployed network functions according to context changes) of VNFs, CNFs, and PNFs in different network domains.

At the same time, the atomization of the classical access-core dichotomy in network domains is paving the road for network *micro-domains* that substantially increase the management granularity of physical infrastructures. Figure 1 shows how distinct

controllers and orchestrators are expected to operate in the following micro-domains: Core, Transport, Edge (running the cloud part of the RAN), and Far Edge (that includes the radio units and the fronthaul).

The radical architectural transformations outlined above summarize the current state-of-the-art in standardization initiatives 3GPP [3] and ETSI [12]. Jointly, these evolutions will enable very fine-grained engineering of network configurations on the fly, so that the network functions and their allocated resources can be adapted to traffic demand dynamics and meet strict, heterogeneous Quality of Services (QoS) specifications. As such, these architectural changes will support network-on-demand schemes, multi-tenancy and emerging compelling paradigms for strong service differentiation such as network slicing that will characterize 6G multi-service settings.

2.2 The role of NI in 6G systems

Fundamental to the optimal operation of the softwarized, cloudified, and atomized network infrastructure will be the Network Intelligence (NI) responsible for managing the composite mosaic of network functions and associated resources in presence of a surging mass of services, tenants, and slices. Given the variety of network management tasks and the many functions they involve, each orchestrator or controller shall run multiple NI instances, each implemented as a pipeline of supple yet effective algorithms that swiftly detect or anticipate new requests or fluctuations in the network activities, and then react to those by instantiating, relocating, or re-configuring network functions in a fully automated manner.

By driving these decisions, NI is expected to meet a number of key targets that include: (*i*) ensuring that all traffic demands are accommodated in a timely fashion and in full conformity to QoS requirements; (*ii*) maximizing infrastructure and resource reuse across multiple tenants or slices to reduce operating expenses; (*iii*) taking full advantage of the novel flexibility to compensate for the reduced performance of equipment where software functions are parted from vendor hardware; and, (*iv*) completely eliminating human intervention, hence fulfilling the vision of zero-touch network and service management in mobile systems [6]. These aspects are all highly critical, to the point that the success and viability of 6G systems will largely depend on the quality and appropriate integration of NI solutions in the network infrastructure.

2.3 Challenges in the integration of NI

Present trends in NI for next-generation network orchestration that are promoted by major standardization bodies pivot on the notion of *closed-loop* Artificial Intelligence (AI) [3, 12]. According to this paradigm, the NI instances deployed at centralized orchestrators and controllers work in closed control loops: abiding by the learning principles of modern AI, they record the context of management decisions, collect observations about the quality of such decisions via continuing monitoring, and then use the feedback to improve future choices. A closed-loop model lets NI apprehend what is important for an operator in a certain situation, and learn over time to automate optimal decision making towards the expected targets (*i*)-(*iv*) listed in Section 2.1.

The current, prevailing vision for closed-loop NI contemplates instances located at centralized orchestrators or controllers in the control plane that interact with Network Functions (NFs) deployed in the data plane [8], exemplified by the local NI loops in Figure 1. This model requires that network state information is gathered from the different NFs and transported to a central entity, where they are processed by the pertinent AI algorithms; decisions must then travel back across the network before they can be enacted. While present standardization efforts and the associated drafts stop at this point, we believe that the architectural changes currently proposed still yield two major limitations, as follows.

First, there is a concrete risk that the latency in data transfer and decision communication ensuing from a strongly centralized closed-loop model hinders the effectiveness of NI, and limits its application in a number of critical network management scenarios. To be effective, many network functionalities must operate at timescales that are not compatible with the lengthy procedures of data gathering at NFs, data processing in the control plane, and final restitution of actions to be implemented in the data plane. In these cases, the feedback loop between the control and data planes does not allow the NI to monitor the system in a continuous manner, and to immediately intervene when required; instead, operators must just hope that the policies, NFs, and resources already in place are sufficient to cope with sudden fluctuations in traffic – or otherwise incur into service disruption and violations of the Service Level Agreements (SLAs) in place with tenants.

Second, the current architectural models do not allow for any interaction among NI instances, which are left to operate in a vacuum, in spite of the fact that the actions they take often yield reciprocal impact. As a simple but representative example, short-timescale network controllers operate on local resources allocated by orchestrators in charge of similar decisions at a wider scope and on longer timescales [5]. Therefore, for NI to operate at its best, it is critical that NI instances can cooperate to ensure end-to-end synchronization, convergence, and global optimality of the zero-touch network management process.

In order to tackle and remove these limitations, we believe that architectural changes to 6G networks must necessarily abide by a *NI-native* model, a concept that we expound next.

3 A NI-NATIVE ARCHITECTURE

Our proposed mobile network architectural model is summarized in Figure 2. Its design anticipates the success of current trends that push automated decisions towards the physical infrastructure, and natively accommodates intelligence beyond the control plane and closer to the user plane. The architectural changes above entail a much deeper integration of intelligence in 6G systems and create a considerably wider ecosystem of NI instances that populate the network infrastructure. The original hierarchy of NI instances removes the restrictions of closed-loop models, as presented in Section 3.1. However, it also creates new needs in terms of coordination, as discussed in Section 3.2.

3.1 NI across network operation timescales

The architecture in Figure 2 outlines three levels of NI, told apart by the timescale of their associated management operations. The three

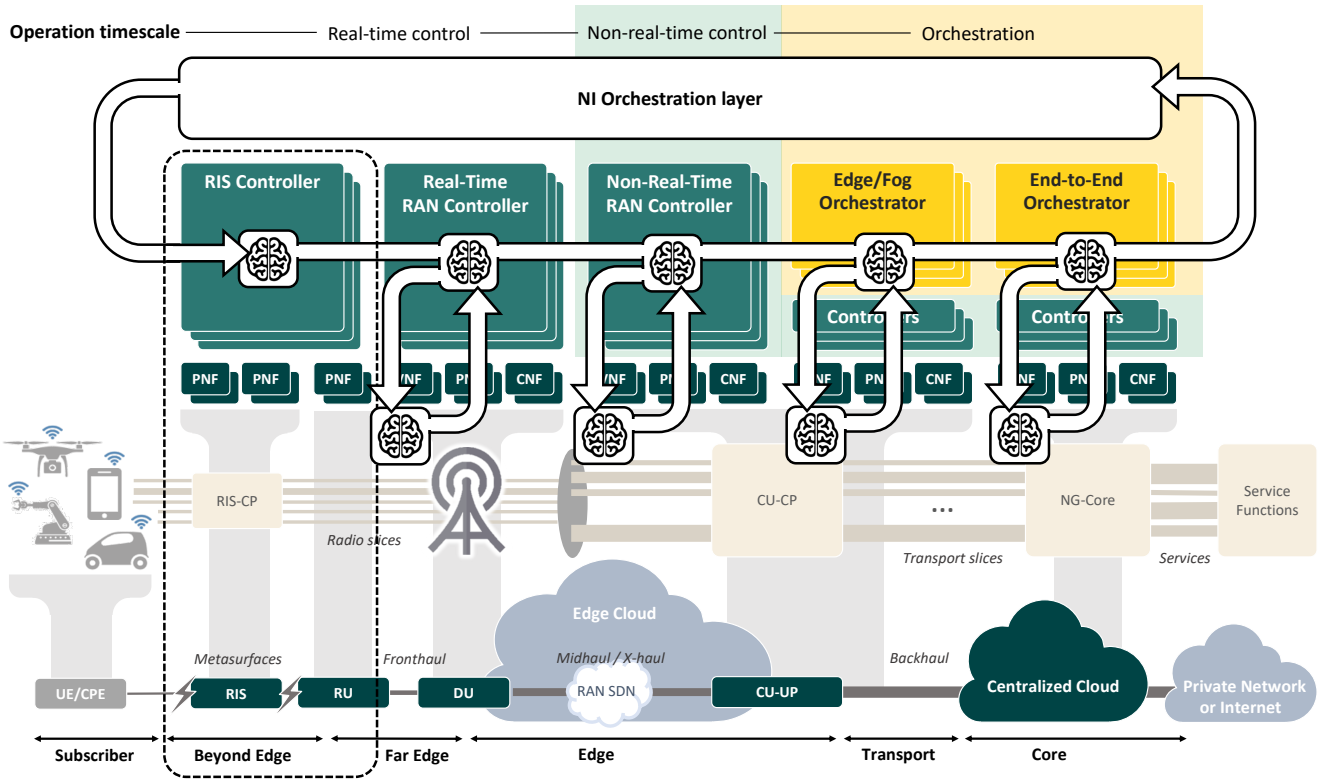

Figure 2: Concept of the proposed NI-native architecture. The three levels of NI operation timescales are set apart by shaded areas that encompass the network entities belonging to each level; the concept brings pervasive intelligence into the real-time control domain of the network infrastructure, all the way within NFs. The additional Beyond Edge micro-domain and its associated entities are highlighted by the dashed line. The novel NI Orchestration layer introduces feedback loops across NI instances deployed throughout the network, including the new ones implemented at the NF level.

levels are highlighted at the top of the figure, and are described as follows.

- **NI at Orchestrators:** in Core Network (CN) deployments, and for service-related applications running in Private Network (PN) deployments, End-to-End Orchestrators operate at a global network level, whereas Edge/Fog Orchestrators focus on broad portions of the mobile edge. The NI integrated into these entities manage functions and resources at a broad network scope, taking decisions that affect large sets of Distributed Units (DUs) and Remote Units (RUs) at once. At this layer, the data traffic is aggregated over a vast population of User Equipment (UE) and its dynamics vary relatively slowly. Reconfiguration is performed at timescales of minutes to hours, and NI must take decisions that endure large time windows.
- **NI at Non-Real-Time Controllers:** shifting network functionalities closer to the user and substantially improving the versatility of Radio Access Networks are long-standing objectives in the evolution of mobile networks, which will finally become a reality with next-generation mobile systems. Here, Non-Real-Time RAN Controllers are responsible for less time-critical functionalities at a limited number of sites. Their associated NI operates at a local scope and takes decisions at timescales in the order of tens of seconds.
- **NI at Real-Time Controllers:** ambitious 6G targets like unperceivable latency or pseudo-infinite bandwidth require data analysis and decision-making at timescales of milliseconds or lower. Standardization activities within O-RAN [1] or FOG-05 [2] expect that the NI deployed within Real-Time RAN Controllers will ensure fast-timescale functionalities such as radio scheduling. However, in contrast to the vision of current standardization bodies, our concept envisions the integration of NI within NFs themselves, across the Far Edge, Edge, Transport, and Core micro-domains. In these cases, NI is implemented much closer to or actually within in the user plane: instead of injecting direct policies into "dumb" VNFs, Controllers trigger highly specialized and lightweight NI algorithms that operate within each NF, enabling, network functionalities to operate close to the line rate. This approach naturally leads to introducing an original *Beyond Edge* micro-domain: there, dedicated Controllers run the NI that actions Reconfigurable Intelligent Surfaces (RIS), which effectively turn a traditionally passive environment where

transceivers react to into a programmable environment that actively assists in the task of delivering data over the air.

The architectural changes above enable very fast and localized decision-making, reducing the reaction times of resource reallocation and VNF reconfiguration in response to fluctuations in the traffic demands and changes of requirements on performance indicators. As a result, it allows surpassing the limitations of closed-loop models highlighted in Section 2.3. However, these NI instances need to interact seamlessly to perform at their best, and exchange data and information so as to mutually improve both their learning and decision-making processes.

3.2 Cross-domain NI orchestration

To address the need for system-wide NI coordination, our NI-native architectural model sets forth a structured approach based on an original *NI Orchestration layer*. It is responsible for supervising intelligence in the network architecture as a whole, ensuring the ideal functioning of each closed-loop NI instance, and overseeing interactions across closed loops that run NI at different timescales. The NI Orchestration layer is devoted to two NI management tasks.

- **NI algorithm selection:** as later explained in Section 4, multiple variants of the same NI shall be designed by employing different modelling strategies and adaptive learning techniques. The NI Orchestrator is then responsible for selecting the best NI algorithm to be run within each NI instance from a predefined range of options. Decisions on the most appropriate option are based on contextual information (*e.g.*, the reliability level of the traffic demand predictions), available system resources (*e.g.*, the computational capacity that can be dedicated to the NI instance), performance requirements (*e.g.*, the target precision of the algorithm), and amount of data to be processed (*e.g.*, the lookback for a forecasting algorithm), concurrently across timescale levels.
- **NI instance coordination:** the NI Orchestrator must ensure the gracious operation of all NI instances operating at different timescales and in different micro-domains. To this end, the NI Orchestrator supports the exchange of information via dedicated interfaces with the NI instances, and centrally solve trade-offs that may emerge from conflicting objectives in the control and data planes, including those associated with establishing policies (at short timescales) versus enforcing such policies (at faster timescales). For instance, this is what happens in the case of radio resource orchestration for bandwidth allocation (policy) versus radio scheduling within such bandwidth (enforcement), but it is in fact the case for many of the NI-assisted functionalities that could be implemented in a mobile network. We remark that our proposed NI-Native architecture goes beyond federated learning aspects, and deals with the practical application of NI in multi-timescale environments where the network metric shall guide the zero-touch management process.

Thus, the NI Orchestration layer establishes and manages a rich and organized set of feedback links between the NI at orchestrators, controllers, and NFs, as highlighted in Figure 2. This allows the NI Orchestrator to help NI algorithms that run in the data plane to overcome the intrinsic limitations of their local view and close in on the optimal point of operation without sacrificing their reactivity. Similarly, the NI Orchestrator supports the NI algorithms in traditional End-to-End Orchestrators by feeding them with fresh local state and decisions at NFs, so that they ensure system-wide stability without human intervention.

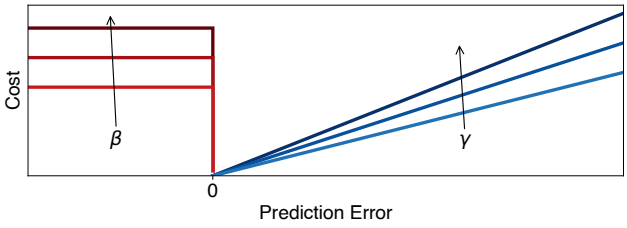

Figure 3: The DeepCog tailored loss function.

4 NI FOR RESOURCE ORCHESTRATION

While the NI-native architecture above creates an ideal environment for the integration of AI in mobile network infrastructures, it is only part of the solution to the problem of imbuing intelligence into 6G systems. Indeed, it must be complemented by a design of AI models that meets the specific requirements of network functionalities, so as to empower NI instances effectively.

In the following, we apply our NI architecture to address resource orchestration. In particular, we develop a capacity forecasting model that anticipates the optimal amount of resources to be allocated to network entities at each reconfiguration opportunity. This algorithm builds on the DeepCog approach we proposed in [5] and is an integral part of the orchestrator operation.

DeepCog addresses the problem of optimizing the utilization of resources across network slices – a vital aspect for the reliable operation of 6G systems characterized by a very high degree of multi-tenancy, where strong service guarantees force isolating resources across slices. Measurement-driven studies show that, already under today's traffic demands, ensuring resource isolation among slices risks to yield unbearable costs for operators, as granting resources to slices under mildly efficient allocation strategies may require a six-fold increase of available capacity [10].

In the above scenario, accurately anticipating the capacity needed to meet the future demand generated by individual slices is paramount to the design of systems that are extremely reliable in satisfying the complex service requirements that 6G systems will face. The notion of capacity is deliberately general, as the problem arises across network domains, where it applies to computing, transport, or radio resources depending on the target entity, and concerns timescales ranging from seconds to hours minutes. Unfortunately, current traffic predictors aim at forecasting traffic demands, and not the capacity needed to serve them, incurring in very frequent underestimation errors that lead to Service Level Agreement (SLA) violations and undermine the reliability of the network [5].

The core of the DeepCog algorithm is the loss function that translates the forecast load level into a feedback signal that proactively steers the network configuration to optimal levels. Indeed, the vast majority of AI models for NI proposed to date employ generic loss functions such as L1, L2, or Cross Entropy, which do not suit the

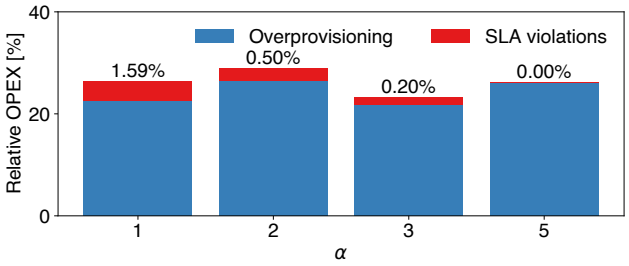

Figure 4: Orchestration cost obtained with DeepCog.

distinctive objectives of each network functionality, hence incur in a *loss-metric mismatch* [9]. The loss-metric mismatch implies that the metrics that shall be optimized, *i.e.*, the quality of service in its different facets, are not adequately captured by the loss function measured on network parameters.

To address the loss-metric mismatch in capacity forecasting, DeepCog employs a customized loss function that is driven by the monetary value resulting from capacity allocation decisions [4]. As illustrated in Figure 3, the loss function associates positive forecast errors to the unnecessary provision of resources, which has a cost that proportionally increases with the amount of their unused quota with a slope γ. Negative errors map instead to insufficient allocated resources, hence service disruption and SLA violations, which yield a high economic penalty β irrespective of the error.

Figure 4 shows the overall OPEX obtained by orchestrating slices with DeepCog for real-life measurements data as a function of $\alpha = \beta/\gamma$, which expresses the cost of one SLA violation relative to that of overprovisioning one unit of capacity [5]. The total OPEX is expressed as the percent extra-economic fee incurred by the operator over an oracle that can perfectly predict future demand, and the number on top of each bar reports the contribution of SLA violation fees. We observe that DeepCog is very effective, as the extra cost over a perfect oracle is only around 20%. Furthermore, DeepCog effectively reduces theSLA violations as their relative cost to overprovisioning (*i.e.*, the α ratio) increases.

5 CONCLUSION

In this paper, we motivated the need for a new NI-native architecture for next-generation (6G) mobile networks. We have further identified the challenges involved in realizing such an architecture. As a first step towards addressing such challenges, we have outlined the key modules and concepts required to bring NI into mobile networks in efficient and effective way. To show the potential benefits of our framework, we have focused on a NI-based resource orchestration solution.

ACKNOWLEDGEMENT

The authors of this paper have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017109 (DAEMON Network intelligence for aDAptive and sElf-Learning MObile Networks).

This paper is also funded by the Spanish State Research Agency (TRUE5G project, PID2019-108713RB-C52PID2019-108713RB-C52 / AEI / 10.13039/501100011033)

REFERENCES

[1] [n.d.]. Eclipse fog05 - The End-to-End Compute, Storage and Networking Virtualisation solution. https://fog05.io/.. Accessed: 2020-10-15.
[2] [n.d.]. O-RAN Alliance. https://www.o-ran.org/.. Accessed: 2020-10-15.
[3] 3GPP. 2017. *Technical Specification Group Services and System Aspects; Study on enablers for network automation for the 5G System (5GS).* Technical Report (TR) 23.701. 3rd Generation Partnership Project (3GPP). Version 17.
[4] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez. 2019. α-OMC: Cost-Aware Deep Learning for Mobile Network Resource Orchestration. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 423–428.
[5] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez. 2020. DeepCog: Optimizing Resource Provisioning in Network Slicing With AI-Based Capacity Forecasting. *IEEE Journal on Selected Areas in Communications* 38, 2 (2020), 361–376.
[6] C. Benzaid and T. Taleb. 2020. AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions. *IEEE Network* 34, 2 (2020), 186–194.
[7] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi. 2020. Toward 6G Networks: Use Cases and Technologies. *IEEE Communications Magazine* 58, 3 (2020), 55–61. https://doi.org/10.1109/MCOM.001.1900411
[8] D. M. Gutierrez-Estevez, M. Gramaglia, A. D. Domenico, G. Dandachi, S. Khatibi, D. Tsolkas, I. Balan, A. Garcia-Saavedra, U. Elzur, and Y. Wang. 2019. Artificial Intelligence for Elastic Management and Orchestration of 5G Networks. *IEEE Wireless Communications* 26, 5 (2019), 134–141.
[9] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. 2019. Addressing the loss-metric mismatch with adaptive loss alignment. In *International Conference on Machine Learning*. PMLR, 2891–2900.
[10] Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Pérez. 2019. Resource Sharing Efficiency in Network Slicing. *IEEE Transactions on Network and Service Management* 16, 3 (2019), 909–923. https://doi.org/10.1109/TNSM.2019.2923265
[11] O-RAN Alliance. 2020. O-RAN-WG1-O-RAN Architecture Description - v02.00.00. Technical Specification.
[12] Y. Wang, R. Forbes, C. Cavigioli, H. Wang, A. Gamelas, A. Wade, J. Strassner, S. Cai, and S. Liu. 2018. Network Management and Orchestration Using Artificial Intelligence: Overview of ETSI ENI. *IEEE Communications Standards Magazine* 2, 4 (2018), 58–65.
[13] C. Zhang, P. Patras, and H. Haddadi. 2019. Deep Learning in Mobile and Wireless Networking: A Survey. *IEEE Communications Surveys Tutorials* 21, 3 (2019), 2224–2287.