Self-Managed 5G Networks

27 July 2024

COMMUNICATIONS NETWORK AND SERVICE MANAGEMENT IN THE ERA OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

COMMUNICATIONS NETWORK AND SERVICE MANAGEMENT IN THE ERA OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

author 1
affiliation

author 2
affiliation

LOGO

Contents

Chapter 1

Self-Managed 5G Networks

1.1. State of the Art

1.1.1. Service orchestration

The evolution towards the 5th generation of cellular mobile networks (5G) consists of dynamically managing and orchestrating several virtual nodes assigned to different and dynamic network slices. The 5G is expected to support a numerous amount of services with versatile quality of service (QoS) demands. Network Function Virtualization (NFV) is a key and promising solution for the service providers towards enabling in an efficient manner the deployment of new type of services and coping with the increasing network traffic and the strict service demands.

Several research works have been proposed in the direction of applying the concept of NFV to 5G networks. Vertical services can indeed be defined as a set of connected virtual network functions (VNFs), often represented as a graph, which can be deployed on the computing, storage, and network resources within the MNO infrastructure. To efficiently support the various vertical services, the network slicing paradigm has also emerged, so that the mobile operator's can deploy different services, ensuring for each of them their isolation requirements as well as their required key performance indices (KPIs), in spite of the limited resources available. Network slicing also supports composed services, i.e., such that the VNF set includes sub-sets, each corresponding to a sub-slice service.

To create a slice, MNOs must decide where to place each VNF and allocate the necessary resources (e.g., virtual machines or containers, and virtual links connecting them). The dynamic placement of VNFs close to the Edge of the network, as well as across the Cloud, Edge, and Fog, have been examined thoroughly, since the optimal selection of a VNF placement solution is known

to be NP-hard. Given the widespread use of linear models, it has become increasingly important to exploit them so as to efficiently support the decision-making in the service orchestration process.

The typical scenario considered for NFV-based 5G networks foresees that the VNF placement decisions are made by NFV Orchestrator (NFVO), defined in the ETSI Management and Orchestration (MANO) framework [3]. However, ETSI [2, Sec. 8.3.6] specifies four granularity levels for placement decisions: individual host, zone (i.e., a set of hosts with common features), zone group, and point-of-presence (PoP) (e.g., a datacenter). Real-world mobile networks implementations, including [7], assume that the NFVO, or similar entities, make PoP-level decisions. Placement and sharing decisions within individual PoPs, instead, can be made by other entities under different names and with slight variations between IETF, the NGMN alliance, and 5G PPP. The latter, in particular, includes a Software-Defined Mobile Network Coordinator (SDM-X), as depicted in Figure 1.1. The SDM-X operates at a lower level of abstraction than the NFVO and makes intra-PoP VNF placement decisions. Specifically, for each service newly requested by a vertical, the SDM-X makes decisions on: (i) whether any subset of the service VNFs can reuse an existing sub-slice; (ii) if not possible, which PoP should host the virtual machine (VM) implementing the VNF; (iii) how to allocate (including scaling up/down) the computational capability of the VMs within the PoP.

As for existing solutions to the VNF placement problem, there is a plethora of scientific papers formulating this problem as an Integer Linear Programming (ILP), considering a variety of KPIs and imposing resource-related constraints in order to meet the targeted performance [23, 9, 6, 13]. More specifically, in [23] Sun et al. propose a time-efficient heuristic offline algorithm which is

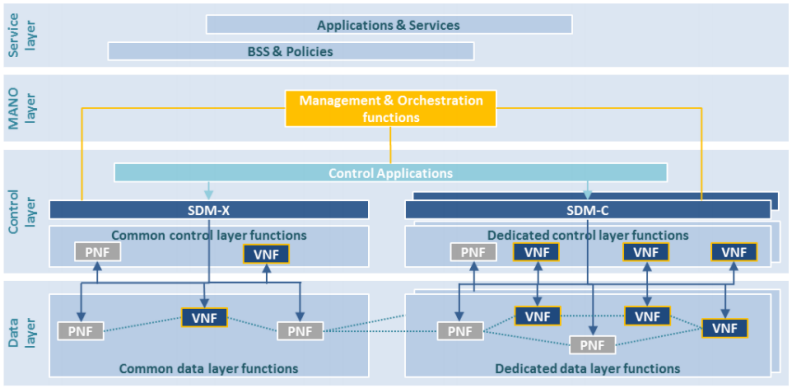

Figure 1.1: 5G-PPP network architecture. Source: [21].

extended so as to predict future VNF demands and reduce the setup delay of the Service Function Chains (SFCs). Davit Harutyunyan et al. in [9] propose an heuristic approach demonstrating a trade-off between the optimality and scalability of the VNF placement solution in large-scale environments. In [6] Richard Cziva et al. employed Optimal Stopping Theory principles to define when to re-evaluate their placement solution, considering the changes in latency values of a real-world scenario and migrate VNFs if necessary. In [13] Luizelli et al. incorporate a Variable Neighborhood Search (VNS), targeting to minimize the number of VNF instances mapped on the Physical Nodes (PNs) and the operational costs respectively.

Since the ILP as the size of mobile network increases becomes more computationally intractable, several papers propose different solutions such as novel complexity-aware heuristic, genetic algorithms and decision tree learning [4, 12, 15, 16]. Importantly, both [14, 16] also address the problem of sub-slices sharing by different service instances, in order to minimize the deployment cost while meeting possible isolation requirements.

Additionally, many recent research works address the VNF placement prob-

lem using Deep Reinforcement Learning (DRL) models due to their efficiency and applicability in many complex problems [17, 5, 11, 18].

Another major research NFV challenge has been the fluctuation of traffic load that needs to be efficiently handled by each VNF instance, so as to meet performance guarantees. Several research works propose novel heuristic and adaptive solutions [10, 24, 20, 8], as well as Neural Network models [22, 19], which assist in proactive auto-scaling decisions by periodically estimating the required number of VNF instances operating on each host PN.

1.2. Future Directions / Some Contributions from the Authors

Admittedly, considering that NFV is still in its early stages, there are numerous open challenges that need to be addressed in order to guarantee the required KPIs, such as ultra-low latency, service availability, and reliability as introduced with 5G. Under the scope of the 5Growth project, novel smart resource orchestration algorithms will be designed and implemented so as to meet the aforementioned requirements.

In particular, it is critical to develop AI/ML approaches that can lead, not only to efficient solutions in terms of KPI fulfillment and resource usage, but also to scalable and flexible mechanisms that can effectively deal with different services and greatly reduce the service deployment time. To this end, and drawing on the work carried out by previous 5G PPP projects, 5GROWTH aims at designing an AI/ML-based architecture where the different aspects of service and resource orchestrations are tackled at different architectural layers.

First, upon receiving a vertical service request, an entity called Vertical Slicer, will take care of (i) assessing whether the service can be accommodated, given the amount of resources the vertical is entitled to use, (ii) which type of slice should be created, (iii) which, if any, sub-slice(s) can be reused for the service deployment. To accomplish these tasks, the Vertical Slicer will exploit data collected through a monitoring platform, query the underlying service orchestrator for information on the resource availability, and use these data to feed an AI/ML model to make sensible decisions.

Equipped with the above decisions, the Service Orchestrator can then perform the actual VNF placement and/or properly scale the resources to be allocated to existing sub-slices so that they can handle the additional traffic load associated with the newly requested service. Importantly, the service orchestrator will have to ensure a smooth functioning of the deployed slices, in spite of time-varying traffic loads and operational conditions. Again, by exploiting monitoring data, further scaling of the resources allocated to a slice may be needed, so as to meet the target KPIs. This may lead to a variation of the amount of resources assigned to the VMs and allocated on the virtual links to dela with a service traffic (scale up/down), or to the creation/deletion of a VNF replica (scale out/in). It is therefore essential to develop AI/ML solutions that can addresse these issues by making near-optimal decisions. On the one hand, reinforcement learning approaches will be adopted for real-time decisions, by defining reward functions that reflect the service requirements. On the other hand, a hybrid algorithm will be implemented, which combines genetic-based solutions with NNs, targeting to minimize end-to-end delay while ensuring that there will be no KPI violations. More precisely, genetic-based algorithms are examined and selected as solutions to the VNF placement problem due to

their fast converge to the near-optimal solution. In this direction, NNs were selected so as to map traffic and application-aware metrics to scaling decisions, POLITO-comment:the following sentence sounds a bit unclear to us. given the fact that in most cases outperform in problems that require pattern recognition in a immense amount of data.

To realize the above vision, it is clear that pre-trained AI/ML models will be needed at the different architectural layers. It is therefore pivotal to develop a platform specifically designed to make off-line trained models available, as well as to provide the Vertical Slicer and the Service Orchestrator with models that are trained on the spot, leveraging the data collected through the monitoring system. This is indeed the core of a closed-loop system as also envisioned by the ETSI Zero- touch network Service Management (ZSM) [**?**]. Open interfaces will have to be developed, which allow service characteristics and target KPIs to be specified by the vertical, passed as hyper-parameters to the Vertical Slicer, and then down to the other layers of the network architecture. In this way, the AI/ML mechanisms can be tailored to the specific service and application requirements, leading to a fully-automated 5G system.

1.2.1. Future Directions / NKUA: Discussion on AI-aware networking (for RAN resource allocation approaches)

The massive computing power that has become available via the worldwide network of data centers offered by ISPs, OTTs, etc., along with the radical progress of the AI and ML sciences themselves, have already started paving

the way for an extensive uptake of AI/ML for diverse network operations. Particularly regarding the RAN resource allocation challenges, in the ultra-dense RAN ecosystems foreseen for 5G - and beyond -, such approaches will often require extreme computing resources, in order to process vast amounts of diverse contextual data. Distributed training, and as well distributed inference techniques are already being proposed in the literature presented earlier, in order to exploit the increased processing capabilities of diverse RAN elements, exploiting even end user device capabilities when applicable. As it can be inferred from the previous paradigms, the AI and ML concepts are gradually becoming structural components of the network, introducing novel capabilities for intelligent RAN resource management, user and network device profiling, spectrum allocation techniques, etc.

Besides the computing resources that are required, one of the most crucial challenges with regard to AI/ML and network operations, is that up to now, AI/ML algorithms and network and communication protocols are designed separately. Joint RAN resource management and AI/ML algorithms design should be one of the high priorities towards a truly *AI-aware networking paradigm*, where coding and signal processing approaches are integrated with the AI framework. This will be realized by identifying common requirements and limitations that result from the two domains; for example, such joint approaches could be considering different dimensionality reduction or data encoding techniques for limited computing capabilities devices (such as IoT nodes), adaptive gradient aggregation for improved resilience, or joint channel coding and image compression techniques in noisy wireless environments.

3GPP has already introduced a novel Network Data Analytics Function (NWDAF) in Rel.16 [1], which indicates the gradual adoption of the ML and

AI concepts in the core network architecture also from the standardization perspective. Currently, this function has limited functionality and is deployed only as part of the Core Network (5GC) in order to facilitate operators' policy manipulation. Additionally, currently the data analytics is limited to only 3GPP-oriented information. Taking the afore-mentioned AI-aware networking paradigm as design guideline, the evolution of such a NF to a distributed, multi-domain, federated learning-based approach, exploiting also resource information from non-3GPP networks, could potentially boost the network AI capabilities towards seamless and more flexible RAN resource management at the Network Edge. On top of that, the extraction of user and network behavioral patterns and their exploitation towards predictive RAN resource allocation, seems to be able of offering considerable additional gains to the current RAN resource management approaches; such an enabler would require radical enhancements in the current architecture and NWDAF operation, enabling a distributed profile extraction approach exploiting edge nodes' - including in several cases even the UEs' - computing power.

Bibliography

[1] *3GPP TS 29.520 version 16.2.0, 5G System; Network Data Analytics Services, Release 16, December 2019.*

[2] ETSI Network Functions Virtualisation (NFV); Management and Orchestration; Or-Vnfm reference point – Interface and Information Model Specification, 2016.

[3] ETSI GS NFV-MAN 001 Network Functions Virtualisation (NFV); Management and Orchestration, Accessed: June 2020. URL https://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_NFV-MAN001v010101p.pdf.

[4] S. Agarwal, F. Malandrino, C. F. Chiasserini, and S. De. Vnf placement and resource allocation for the support of vertical services in 5g networks. *IEEE/ACM Transactions on Networking*, 27(1):433–446, 2019.

[5] H. Chai, J. Zhang, Z. Wang, J. Shi, and T. Huang. A parallel placement approach for service function chain using deep reinforcement learning. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pages 2123–2128, 2019.

[6] R. Cziva, C. Anagnostopoulos, and D. P. Pezaros. Dynamic, latency-optimal vnf placement at the network edge. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 693–701, 2018.

[7] Antonio De la Oliva, Xi Li, Xavier Costa-Perez, Carlos J Bernardos, Philippe Bertin, Paola Iovanna, Thomas Deiss, Josep Mangues, Alain Mourad, Claudio Casetti, et al. 5g-transformer: Slicing and orchestrating transport networks for industry verticals. *IEEE Communications Magazine*, 2018.

[8] X. Fei, F. Liu, H. Xu, and H. Jin. Adaptive vnf scaling and flow routing with proactive demand prediction. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 486–494, 2018.

[9] D. Harutyunyan, N. Shahriar, R. Boutaba, and R. Riggio. Latency-aware service function chain placement in 5g mobile networks. In *2019 IEEE Conference on Network Softwarization (NetSoft)*, pages 133–141, 2019.

[10] O. Houidi, O. Soualah, W. Louati, M. Mechtri, D. Zeghlache, and F. Kamoun. An efficient algorithm for virtual network function scaling. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–7, 2017.

[11] H. R. Khezri, P. A. Moghadam, M. K. Farshbafan, V. Shah-Mansouri, H. Kebriaei, and D. Niyato. Deep reinforcement learning for dynamic reliability aware nfv-based service provisioning. pages 1–6, 2019.

[12] M. A. Khoshkholghi, J. Taheri, D. Bhamare, and A. Kassler. Optimized service chain placement using genetic algorithm. In *2019 IEEE Conference on Network Softwarization (NetSoft)*, pages 472–479, 2019.

[13] Marcelo Caggiani Luizelli, Weverton Luis, Luciana S. Buriol, and Luciano Paschoal Gaspary. A fix-and-optimize approach for efficient and

large scale virtual network function placement and chaining. *Computer Communications*, 102:67 – 77, 2017.

[14] F. Malandrino, C. F. Chiasserini, G. Einziger, and G. Scalosub. Reducing service deployment cost through vnf sharing. *IEEE/ACM Transactions on Networking*, 27(6):2363–2376, 2019.

[15] D. M. Manias, M. Jammal, H. Hawilo, A. Shami, P. Heidari, A. Larabi, and R. Brunner. Machine learning for performance-aware virtual network function placement. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2019.

[16] J. Martinéz-Peréz, F. Malandrino, C.F. Chiasserini, and C.J. Bernardos. OKpi: All-KPI network slicing through efficient resource allocation. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020.

[17] J. Pei, P. Hong, M. Pan, J. Liu, and J. Zhou. Optimal vnf placement via deep reinforcement learning in sdn/nfv-enabled networks. *IEEE Journal on Selected Areas in Communications*, 38(2):263–278, 2020.

[18] P. T. A. Quang, Y. Hadjadj-Aoul, and A. Outtagarts. A deep reinforcement learning approach for vnf forwarding graph embedding. *IEEE Transactions on Network and Service Management*, 16(4):1318–1331, 2019.

[19] S. Rahman, T. Ahmed, M. Huynh, M. Tornatore, and B. Mukherjee. Auto-scaling vnfs using machine learning to improve qos and reduce cost. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018.

[20] Y. Ren, T. Phung-Duc, Y. Liu, J. Chen, and Y. Lin. Asa: Adaptive vnf scaling algorithm for 5g mobile networks. In *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, pages 1–4, 2018.

[21] Peter Rost, Christian Mannweiler, Diomidis S Michalopoulos, Cinzia Sartori, Vincenzo Sciancalepore, Nishanth Sastry, Oliver Holland, Shreya Tayade, Bin Han, Dario Bega, et al. Network slicing to enable scalability and flexibility in 5G mobile networks. *IEEE Comm. Mag.*, 2017.

[22] Tejas Subramanya, Davit Harutyunyan, and Roberto Riggio. Machine learning-driven service function chain placement and scaling in mec-enabled 5g networks. *Computer Networks*, 166:106980, 2020.

[23] Q. Sun, P. Lu, W. Lu, and Z. Zhu. Forecast-assisted nfv service chain deployment based on affiliation-aware vnf placement. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2016.

[24] Adel Nadjaran Toosi, Jungmin Son, Qinghua Chi, and Rajkumar Buyya. Elasticsfc: Auto-scaling techniques for elastic service function chaining in network functions virtualization-based clouds. *Journal of Systems and Software*, 152:108 – 119, 2019.