Processing Modular Application Functions in Future Medical 6G Radio Access Networks

Nicolai Kröger[1], Sven Kolb[2], Franziska Jurosch[2], Dirk Wilhelm[2,3,4], Wolfgang Kellerer[1]
[1]*Technical University of Munich, TUM School of Computation, Information and Technology,*
Chair of Communication Networks
[2]*Technical University of Munich, TUM School of Medicine and Health, Klinikum rechts der Isar, Research Group MITI*
[3]*Technical University of Munich, TUM School of Medicine and Health, Klinikum rechts der Isar, Department of Surgery*
[4]*Technical University of Munich, Munich Institute of Robotics and Machine Intelligence,*
firstname.lastname@tum.de

***Abstract*—The currently researched 6G communication standard promises to enable new applications with the native integration of new technologies and concepts. Especially the medical area will greatly benefit from the developments of new communication networks. As the communication infrastructure will be able to satisfy stringent and varying requirements, emerging medical applications such as telemedicine and telesurgery will be enabled. However, the concrete networking architecture remains an open question for research. In this paper, we propose a dedicated medical 6G Radio Access Network (RAN) architecture and discuss possible usage scenarios based on two medical use cases. First, a network architecture is presented which leverages in-network computing to execute medical applications. The main idea of this concept is the dynamic interaction between medical applications and the network. In particular, the placement of Modular Application Functions (MAFs) and their execution depends on the state of the network. Secondly, we map two medical use cases, namely a semiautonomous telerobotic examination suite and a context-sensitive medical environment, to the proposed network architecture and explain the interaction between our network architecture and these applications in detail. Our approach demonstrates the potential of the combined development of medical applications and their underlying communication architecture.**

***Index Terms*—6G, In-Network Computing, Virtualization, MAF, Medical Technology.**

I. INTRODUCTION

The fifth generation of mobile communication standards (5G) aimed to satisfy the requirements of each possible application. The goal was to provide a flexible underlying communication infrastructure capable of serving every demand. However, since the roll-out of 5G, its limitations have become more and more visible. In particular, emerging applications such as virtual reality (VR) and connected autonomous systems place demands on the networking infrastructure which 5G cannot satisfy [1]. Thus, research toward 6G is currently trying to address those challenges. With 6G, the performance of networking parameters such as throughput, latency and availability is envisioned to increase tremendously. To achieve this, 6G follows a holistic design which fully integrates applications within the network, introducing new technologies such as higher frequencies, as well as new concepts such as the native integration of Machine Learning (ML) and Artificial Intelligence (AI) [1], [2].

With these new features, 6G will enable new applications, which will particularly suit the medical field [3]. Medical applications are challenging, since they can place stringent requirements on the underlying communication infrastructure with potentially severe consequences in case of failure to satisfy them. Thus, in-network computing is leveraged in 6G to process applications [2]. The challenge lies now in the optimal placement of such medical applications within the in-network computing resources. Existing work has already tackled this challenge in a comparable area. For example, the authors in [4], [5], and [6] investigate the placement of Virtual Network Functions (VNFs) on processing nodes in existing networks. Furthermore, Hentati *et al.* [7] optimize the VNF placement in a medical telerobotic setup. However, with the emergence of 6G and therefore the introduction of new technologies and concepts, the placement of applications and their integration within the network must be re-evaluated. Especially demands which origin from different use cases and therefore with various requirements, typical for medical use cases, have to be addressed.

Thus, in this paper we introduce and discuss a new architecture combining 6G networking infrastructure and application placement within the network. In particular, we combine the design of applications and the design of the 6G network. For this, we abstract applications (or parts of them) as Modular Application Functions (MAFs). Those MAFs are then placed on in-network computing resources. In this way, new processing capabilities are added to the network, leveraging Software-Defined Networking (SDN) and programmable networking devices. The main advantage of using MAFs lies in their flexible description with various parameters and the close interaction with the network, allowing it to change the occupation of networking resources within an application and therefore to use different levels of service, depending on the state of the

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the program of “Souverän. Digital. Vernetzt.”. Joint project 6G-life, project identification number: 16KISK002.

network. In a second step, we map two medical use cases, namely a semiautonomous telerobotic examination suite for remote patient examination and a context-sensitive medical environment for clinical patient monitoring, to the proposed concept and explain in detail how our approach benefits such medical scenarios. The main message of this paper is that the combined development of MAFs and the underlying 6G network shows a high potential to increase the overall achieved performance. In detail, the main contributions of this paper are:

- We introduce an architecture for new 6G networks with in-network computing capabilities to run MAFs. The MAFs interact closely with the network by adapting their service levels to the available networking resources.
- We apply our architecture to two medical use cases and discuss them in detail.

The remainder of the paper is structured as follows. In Section II, background and related work are provided. Next, we introduce our concept for future 6G networks in Section III. In Section IV, we describe two medical use cases and map them to our proposed architecture. Finally, Section V concludes the paper.

II. BACKGROUND AND RELATED WORK

In this section, we provide background and related work on the existing concepts such as networking and placement concepts on which we base our approach.

A. Software-Defined Networking and Programmable Data Planes

In the paradigm of SDN [8], the control plane of networking devices such as switches is separated from the packet-forwarding data plane. This approach enables a network operator to control several switches with one controller and therefore enables network optimization by leveraging a more complete overview of the network. Furthermore, it enables the straightforward deployment of devices from different vendors. One famous protocol implementation of SDN is *OpenFlow* [9]. OpenFlow provides a common interface, allowing the controller to insert rules into the match-action tables of the switches. Those rules consist of a key to which a packet is matched, and a corresponding action which is executed upon matching. Since these actions can only be taken from a predefined set of actions, Programmable Data Planes (PDPs) have been introduced in a next step. PDPs allow the network operator to precisely define the way a packet is handled in a switch. One well-known implementation of combined SDN and PDP is the domain-specific programming language P4 [8]. Extending the features of OpenFlow, the network operator is able to define the packet header parsing in P4, allowing the easy and fast implementation of custom protocols. Furthermore, the executed actions upon matching can now be defined in detail. However, despite the aim of P4 to achieve target-independence, this is not generally valid in practice. The reason for this lies in the actions and the capabilities of switches. If a P4 switch does not support some defined action of a P4 program, the P4 program needs to be adapted and cannot be ported from another target.

B. In-network Computing and Edge Computing

The computing of applications in our proposed approach is based on the combination of various existing computing concepts to execute applications such as *in-network computing* and *edge computing*. Thus, it is worth to take a closer look at existing definitions for a better distinction. The authors in [10] provide an exhaustive survey on this topic. According to them, edge computing (and various related concepts such as fog computing, mobile edge computing, etc.) was introduced in order to offload traffic processing in data centers and reduce latency by placing processing capabilities closer to the end devices. Since the network traffic is increasing further and further, the concept of *in-network computing* was introduced in a next step. As there is no standard definition of in-network computing, the authors in [10] compare different definitions in literature. In principle, they describe in-network computing as traffic processing on networking elements. However, the integration of edge computing within the controlling of the network and the close interaction between pure networking devices and general-purpose processing capabilities such as SmartNICs and CPU-based processing units blurs the border between edge computing and pure in-network computing on switches. Thus, in this paper we use in-network computing for every sort of computation within the network, since the term highly emphasizes the close interaction between the network and computing resources.

C. Placement of Virtual Network Functions

Literature is already rich on different solutions for function placement on in-network processing nodes. In particular, the very similar problem of optimally placing Virtual Network Functions (VNFs) with different aspects has been subject to exhaustive research. Harutyunyan *et al.* investigate the placement of VNFs and their resource allocation, optimizing latency, service costs and migration frequency. VNF migration with minimized power consumption is evaluated in [11]. By optimizing costs and resource consumption in IoT networks, the authors in [12] present a solution for the VNF reconfiguration problem. Focusing on minimizing the deployment costs of radio resource allocation and VNF placement, [13] proposes a new placement approach which considers end-to-end latency. Further optimization of resource usage constrained by the end-to-end latency is done by the authors of [14] and [15] with the usage of parallel executed and shared VNFs. In [16], the authors consider variable usage of computational resources by VNFs and dynamically resize them to optimize resource usage. Varasteh *et al.* introduce a heuristic framework for VNF placement and routing with power consumption and resource constraints. Applying VNF placement in a medical scenario, Hentati *et al.* formulate a joint VNF placement and scheduling problem, optimizing deployment cost with respect to various parameters such as maximum latency, to conduct robotic surgery. Thiruvasagam *et al.* focuses more on the

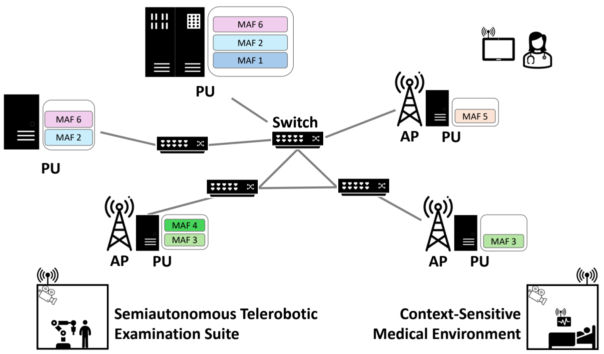

Fig. 1. Network architecture of the envisioned 6G RAN consisting of Access Points (APs), (programmable) switches and Processing Units (PUs), executing Modular Application Functions (MAFs) of the two considered medical use cases.

optimized availability when placing VNFs in 5G networks. Also targeting 5G networks, the authors in [17] introduce a VNF placement and traffic prediction framework for 5G O-RAN and the authors in [18] consider the dynamic VNF and traffic deployment for more real-world parameters. In [6], the authors target future 6G networks with a latency-sensitive VNF placement and scheduling optimization problem considering VNF reuse. Furthermore, VNF placement frameworks leveraging Machine Learning (ML) and Artificial Intelligence (AI) have been introduced by [19], [20].

While the previous work already elaborates on the placement of VNFs considering various parameters, in this work we extend those approaches, leveraging existing concepts such as SDN and PDP, as well as the native integration of in-network computing resources, to form a new 6G network concept which is specifically designed for (but not limited to) medical application scenarios. In particular, we explain such a network in detail and introduce a new form of virtual function: the MAF. In contrast to VNFs, MAFs also abstract applications or parts of applications. Using this approach, the network and applications executed within the network can interact even closer and react to changing situations, allowing for an overall better networking performance and end-user experience.

III. Envisioned 6G RAN Concept

In this section, we present our concept for the envisioned 6G RAN network as shown in Fig. 1. In the following subsections, we first explain the network components in detail and then introduce the concept of MAFs.

A. Network Components

As presented in Fig. 1, the network mainly consists of Access Points (APs), Processing Units (PUs) and switches. *APs* abstract the access technology of the user device, called User Equipment (UE), to the network. The underlying technology such as LiFi, WiFi, 5G or 6G base stations depends on the scenario. Considering a smart medical operating room (OR), a UE could be, for example, a laparoscopic camera. It might be connected to the 6G network via a 6G base station utilizing high frequencies for a data rate up to 1 Tbit/s [1]. Other possibilities are the connection via LiFi for high data rates and due to security reasons, as light cannot exit a closed room. For optimal resource usage and performance, a combination of LiFi and WiFi also is possible, as proposed in [21]. Note that although wired connections are still possible, we focus only on wireless connections in this paper.

PUs abstract the processing capabilities within the network. For that, each PU is described by a certain set of attributes. Examples of such attributes might be overall CPU/GPU computation capability, current utilization, power consumption, memory, possible throughput, current traffic throughput, etc. Furthermore, they can be characterized by the distance to an AP and therefore the latency to the UE. Additionally, each PU might have certain special features such as hardware acceleration or specific CPU generations. The advantage of such a description is that every possible processing capability within the network can be described and used for the placement of MAFs, including conventional PCs, data centers, SmartNICs, etc. In this case, the physical deployment of such a PU is of interest. Similar to the DU/CU concept in 5G [22], smaller sized PUs (in terms of processing capability) are usually located closer to the UE, providing lower latency, whereas larger ones might be integrated as data centers located further away, introducing a higher delay. In a medical OR, smaller PUs could be placed on-site, whereas larger PUs are located off-site for more complex processing such as ML applications.

The last components in the network, *switches*, connect all components within the network. They are SDN-enabled and programmable with domain-specific programming languages such as P4 [8]. As a network operator can now specify the packet handling in detail, this enables the use of switches to perform very fast processing actions on each packet. An idea of this fast processing for already existing P4 devices is presented in [23]–[25], where the authors analyze and model P4 processing for various targets and programs. Furthermore, custom protocols can be implemented. Since such switches provide processing capabilities with their programmability, they can also be abstracted as PUs. However, those processing capabilities are usually limited compared to general-purpose processors, since they depend on the actual implementation of the devices and their programming language with its own limitations [26]. In a clinical scenario, such switches could be used for very fast responses, e.g. to abnormal vital parameters. Furthermore, they enable to duplicate packet streams for critical applications such as telesurgery with several experts at different locations.

B. Modular Application Functions

MAFs are the core of our new concept. They abstract all sorts of applications which can be executed within the network. This includes network functions necessary to run the network such as SDN controllers or 6G radio resource allocation. It further includes applications such as the algorithms applied within the medical use cases described in this paper. Similar to PUs, MAFs can be described with different requirements. They can include (but are not limited to) throughput, latency, availability, priority, CPU consumption, memory consumption, etc. Furthermore, each MAF has certain requirements for the PUs on which it can be executed. As an example, P4 programs can only be executed on P4-enabled networking devices. The exact requirements for the MAFs may vary for specific scenarios and is currently still subject to research. In Section IV, we aim to provide first insights for the design of MAFs in a medical context.

The MAF concept seems close to the Application Function (AF) concept in 5G. However, with our approach we capture all sorts of executable functions. Additionally, MAFs interact more closely with the network resources. The concept of MAFs not only allows the placement of applications in the network with respect to certain constraints. Rather, MAFs are designed to interact with the network, opening up the possibility to adapt applications to network conditions by changing their requirements on the network and networking resources. More precisely, this is achieved by assigning certain requirements to MAFs, such as maximum tolerable latency and desired latency or minimum tolerable throughput and desired throughput. These serve as an upper and lower bound for performance requirements. An application designer creating a new MAF can now consider this during development and adapt the MAF's behavior according to provided networking resources. For example, if the network can no longer provide sufficient resources to satisfy all demanded networking resources in terms of throughput, an MAF streaming video data from a camera could reduce the video resolution within the range of the throughput requirement. Thus, MAFs interact closely with the network, featuring a reactive behavior and achieving an optimized performance of both the network and the application.

In order to further improve the performance in the network, MAF placement is envisioned to be dynamic. This means that – depending on the network state – MAFs can, once running, be dynamically reallocated to another PU for further performance optimization. Similar to the VNF placement as in [4] and [11], this flexibility introduces further challenges for the network and MAF design. Availability is an important factor especially for critical applications; therefore solid approaches for reallocation must be designed for both MAFs and the network. Additionally, the migration of stateful MAFs, i.e., MAFs using memory to store data, needs to be investigated as well as the sharing of MAFs for applications with different requirements.

Another interesting aspect of placing MAFs is the method used for placement. Typically, the placement problem is formulated as an optimization problem such as in [7] where an objective such as minimizing placement costs, maximizing throughput, etc. is optimized with respect to certain constraints such as maximum latency or minimal availability. However, the time needed to run such an optimization problem and find an optimal solution is usually very high, especially if many parameters are used. This long execution time may be problematic for real-time usage with critical applications. Therefore, heuristics are required to reach a near-optimal solution while only using a fraction of the execution time compared to finding an optimal solution. Another approach to place MAFs lies in the native integration of ML and AI in 6G networks. However, compared to the analytical solution-finding approach, problems might occur if situations are different than what the ML/AI algorithm is trained for. This can pose a huge risk for critical applications such as those in a medical context and potentially threaten the well-being of humans. Thus, the employed placement approach needs to be evaluated and adapted for each use case.

The proposed concept of MAFs running within future 6G networks shows great potential to improve the overall performance and enable emerging applications. However, there are still several challenges which need to be addressed. In the next section, we investigate our concept in a medical environment. For this purpose, we describe two medical use cases in detail and explain how they leverage our proposed networking and MAF concept.

IV. MEDICAL USE CASES AND 6G RAN

In the following, we discuss the proposed 6G network executing MAFs, as introduced in Section III, based on two medical use cases, namely a semiautonomous telerobotic examination suite and a context-sensitive medical environment.

A. Semiautonomous Telerobotic Examination Suite

High-precision telemedical applications are being developed to reduce the healthcare gap between metropolitan and rural areas. For this purpose, the usage of robotic telemedical systems is especially promising, since it can enable the provision of comprehensive remote medical examinations by integrating video consultations with robot-assisted diagnostics [3], [27]–[30]. The increasing complexity of telemedical use cases, especially when utilizing human-robot interaction scenarios, correlates with increasingly stringent demands on the underlying communication infrastructure. As an example, the semiautonomous telerobotic examination suite developed in [31] uses synchronous transmission of multiple high-resolution video streams and high-quality multimodal sensor data, along with ultra-reliable transmission of control commands to enable safe teleoperation of a robotic telediagnostic system. These features place large demands on the latency, throughput and reliability of the network [31]–[34].

The semiautonomous telerobotic examination suite leverages the MAF concept to improve the system performance and reduce the hardware footprint of local UE by enabling the placement of computationally intensive application components on network resources. A possible system architecture using the MAF concept based on [31] is illustrated in Fig. 2, where functions involving robotic control (MAF 1-5), video conferencing (MAF 6-8) and medical device interaction (MAF 9-10) are combined into a multimodal distributed application suite. In this example, the computer vision (CV) component

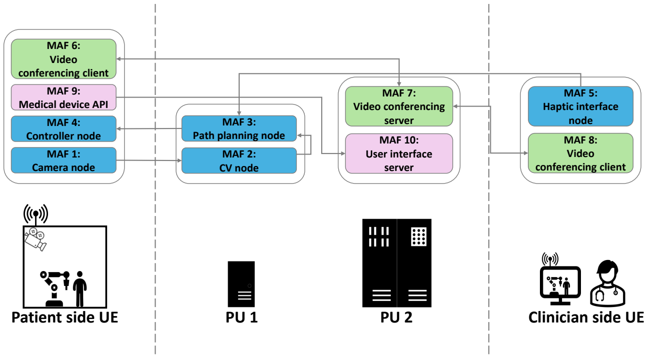

Fig. 2. Possible MAF-based system architecture for a telerobotic examination suite based on [31]. Blue: Robotic control pipeline (MAF 1-5). Green: Video conferencing pipeline (MAF 6-8). Pink: Medical device pipeline (MAF 9-10).

(MAF 2) of the robotic control pipeline places a considerable, constant load on the system resources, making it a suitable candidate for in-network placement. Likewise, the system performance can further be optimized by offloading functions with a smaller computational burden and low latency requirements, such as the path planning node (MAF 3), allowing them to be placed anywhere in the network. In [31], we implement such a system exemplarily and discuss various MAF placement strategies.

B. Context-Sensitive Medical Environment

The goal of context-sensitive applications is to combine process-specific knowledge with real-time environmental information to enable systems to react and respond to changing situations in dynamic environments. In the medical field, the integration of such applications and their augmentation with patient information, medical process models and other situational data can provide an important framework for developing new assistive technologies to reduce the workload of medical professionals and improve patient care. As described in [35], the use of context-sensitive information to trigger actions in medical applications can enable the autonomous adaptation of such systems in highly dynamic hospital environments. Another promising research area is the development of AI-based applications to enhance available clinical information or perform autonomous tasks in healthcare settings [36]. As an example, the situation-dependent adjustment of patient monitoring devices can help to deliver necessary information at the right time, depending on the current clinical context. However, the increased use of such computationally intensive applications correlates with a steady increase in computing resources in the hospital [37]. Developing new context-sensitive applications using the MAF concept can help to alleviate these demands by enabling the use of network resources [38], [39].

C. Prioritization in Medical Applications

A prioritization of application components, data streams and – in the concept proposed in this paper – MAFs, can be considered by classifying them on an application-specific level, as demonstrated in [40]. In medical applications, this can be facilitated to a certain degree by ranking the criticality of MAFs or information streams for the intended medical use case. In scenarios where the satisfaction of application requirements cannot be guaranteed by the underlying network, the prioritization of MAFs can positively impact the reliability of medical applications, thus helping to facilitate efficient medical workflows and improve patient safety. Considering the semiautonomous telerobotic examination suite described above, it is possible to rank the different data streams by priority. For example, the image stream from the patient-side camera to the clinician should have a higher priority than the stream from the clinician-side camera, since the visual monitoring of the patient side is critical not only for diagnostic purposes, but also to ensure the safety of the patient during the examination. Likewise, manual control commands to facilitate teleoperation from the clinician-side haptic interface to the robotic system must be prioritized over the transmission of peripheral data to an external consulting physician. Regarding MAFs, a path planning module placed on network resources must receive a higher priority than e.g. a status logging function, since the former is essential for the robotic control pipeline. Considering the context-sensitive medical environment, MAFs which are essential to a clinical process workflow, such as vital parameter monitoring functions, must be prioritized over other assistive functions such as AI tracking algorithms. In this use case, it may also be beneficial to prioritize certain data streams, such as information from specific vital parameter sensors, over others depending on their relevance to the current medical situation. Finally, optimizing resource utilization by sharing certain MAFs (e.g. vital parameter monitoring functions) among multiple applications with specific requirements and priorities is another idea worth exploring.

V. CONCLUSION

In this paper, we propose a new 6G RAN concept for medical applications leveraging in-network computing. For this, we abstract medical applications (or parts of them) as MAFs with various parameters and place them on processing capabilities within the network. The network interacts with the MAFs and can dynamically adjust the provided level of service in terms of networking resources, according to changes in the network. Based on two medical use cases, a semiautonomous telerobotic examination suite and a context-sensitive medical environment, we study a potential use of medical applications within such a network architecture in detail. Our study shows the potential of the combined development of medical applications and the underlying networking architecture.

In the future, we will optimize the placement of MAFs in such a 6G network for various objectives such as minimized costs or a maximized number of executed MAFs and implement both use cases to demonstrate the capabilities of our concept.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, 2020.

[2] M. Hoffmann and et al., "A secure and resilient 6G architecture vision of the german flagship project 6G-ANNA," *IEEE Access*, vol. 11, 2023.

[3] S. Kolb, F. Jurosch, N. Kröger, F. Mehmeti, L. Bernhard, J. Fuchtmann, S. Speidel, W. Kellerer, and D. Wilhelm, "6G in clinical applications: Integrating new network approaches in healthcare," in *Proc. of CURAC*, 2023.

[4] D. Harutyunyan, N. Shahriar, R. Boutaba, and R. Riggio, "Latency-aware service function chain placement in 5G mobile networks," in *Proc. of IEEE NetSoft*, 2019.

[5] P. K. Thiruvasagam, A. Chakraborty, A. Mathew, and C. S. R. Murthy, "Reliable placement of service function chains and virtual monitoring functions with minimal cost in softwarized 5G networks," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, 2021.

[6] N. Promwongsa, A. Ebrahimzadeh, R. H. Glitho, and N. Crespi, "Joint VNF placement and scheduling for latency-sensitive services," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 4, 2022.

[7] A. Hentati, A. Ebrahimzadeh, R. H. Glitho, F. Belqasmi, and R. Mizouni, "Remote robotic surgery: Joint placement and scheduling of VNF-FGs," in *Proc. of IEEE CNSM*, 2022.

[8] P. Bosshart and et al., "P4: programming protocol-independent packet processors," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, 2014.

[9] N. McKeown and et al., "OpenFlow: enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, 2008.

[10] S. Kianpisheh and T. Taleb, "A survey on in-network computing: Programmable data plane and technology specific applications," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, 2023.

[11] V. Eramo, M. Ammar, and F. G. Lavacca, "Migration energy aware reconfigurations of virtual network function instances in NFV architectures," *IEEE Access*, vol. 5, pp. 4927–4938, 2017.

[12] Y. Liu, Y. Lu, X. Li, Z. Yao, and D. Zhao, "On dynamic service function chain reconfiguration in IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 11, 2020.

[13] N. Gholipoor, H. Saeedi, N. Mokari, and E. A. Jorswieck, "E2E QoS guarantee for the tactile internet via joint NFV and radio resource allocation," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, 2020.

[14] F. Malandrino, C. F. Chiasserini, G. Einziger, and G. Scalosub, "Reducing service deployment cost through VNF sharing," *IEEE/ACM Transactions on Networking*, vol. 27, no. 6, 2019.

[15] S. Kianpisheh and R. H. Glitho, "Joint admission control and resource allocation with parallel VNF processing for time-constrained chains of virtual network functions," *IEEE Access*, vol. 9, 2021.

[16] K. Akahoshi, F. He, and E. Oki, "Service deployment model with virtual network function resizing," *Proc. of IEEE GLOBECOM*, 2021.

[17] K. Ali and M. Jammal, "Proactive VNF scaling and placement in 5G O-RAN using ML," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, 2024.

[18] M. Golkarifard, C. F. Chiasserini, F. Malandrino, and A. Movaghar, "Dynamic VNF placement, resource allocation and traffic routing in 5G," *Computer Networks*, vol. 188, 2021.

[19] V. Eramo and T. Catena, "Application of an innovative convolutional/LSTM neural network for computing resource allocation in NFV network architectures," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, 2022.

[20] J. Chen, X. Cheng, J. Chen, and H. Zhang, "A lightweight SFC embedding framework in SDN/NFV-enabled wireless network based on reinforcement learning," *IEEE Systems Journal*, vol. 16, no. 3, 2022.

[21] H. Vijayaraghavan, J. von Mankowski, and W. Kellerer, "ComputiFi: Latency-optimized task offloading in multipath multihop lifi-wifi networks," *IEEE Open Journal of the Communications Society*, 2024.

[22] A. Martínez Alba and W. Kellerer, "Dynamic functional split adaptation in next-generation radio access networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, 2022.

[23] N. Kröger, H. Harkous, F. Mehmeti, and W. Kellerer, "Performance modeling and analysis of P4 programmable devices with general service times," *IEEE Transactions on Network and Service Management*, 2024.

[24] N. Kröger, F. Mehmeti, H. Harkous, and W. Kellerer, "Performance analysis of general P4 forwarding devices with controller feedback: Single- and multi-data plane cases," *Computer Communications*, vol. 209, 2023.

[25] D. Franco and et al., "A comprehensive latency profiling study of the tofino P4 programmable ASIC-based hardware," *Computer Communications*, vol. 218, 2024.

[26] H. Harkous, C. Papagianni, K. De Schepper, M. Jarschel, M. Dimolianis, and R. Pries, "Virtual queues for P4: A poor man's programmable traffic manager," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, 2021.

[27] A. Asiri, S. AlBishi, W. AlMadani, A. ElMetwally, and M. Househ, "The use of telemedicine in surgical care: A systematic review," *Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH*, vol. 26, no. 3, 2018.

[28] F. Recker, E. Höhne, D. Damjanovic, and V. S. Schäfer, "Ultrasound in telemedicine: A brief overview," *Applied Sciences*, vol. 12, no. 3, 2022.

[29] C. Evans, M. Medina, and A. Dwyer, "Telemedicine and telerobotics: from science fiction to reality," *Updates in Surgery*, vol. 70, 2018.

[30] J. Fuchtmann and et al., "New Method for Surgical Diagnostics - a Robotic Telemedical Approach," *Surgical technology international*, vol. 39, 2021.

[31] S. Kolb and et al., "6G in medical robotics: Development of network allocation strategies for a telerobotic examination system," *International Journal of Computer Assisted Radiology and Surgery*, 2024.

[32] P. Krejov and A. Grunnet-Jepsen, "Intel RealSense Depth Camera over Ethernet," https://dev.intelrealsense.com/docs/depth-camera-over-ethernet-whitepaper [Accessed: (10/04/2024)].

[33] P. Barba, J. Stramiello, E. K. Funk, F. Richter, M. C. Yip, and R. K. Orosco, "Remote telesurgery in humans: A systematic review," *Surgical endoscopy*, vol. 36, no. 5, 2022.

[34] A. Kumcu and et al., "Effect of video lag on laparoscopic surgery: Correlation between performance and usability at low latencies," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 13, no. 2, 2017.

[35] M. Kasparick, M. Schmitz, B. Andersen, M. Rockstroh, S. Franke, S. Schlichting, F. Golatowski, and D. Timmermann, "OR. NET: a service-oriented architecture for safe and dynamic medical device interoperability," *Biomedical Engineering/Biomedizinische Technik*, vol. 63, no. 1, 2018.

[36] F. Jurosch, L. Wagner, A. Jell, E. Islertas, D. Wilhelm, and M. Berlet, "Extra-abdominal trocar and instrument detection for enhanced surgical workflow understanding," *International Journal of Computer Assisted Radiology and Surgery*, 2024.

[37] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Information Fusion*, vol. 66, 2021.

[38] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future healthcare journal*, vol. 8, no. 2, 2021.

[39] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, 2019.

[40] R. Annur, N. Wattanamongkhol, S. Nakpeerayuth, L. Wuttisittikulkij, and J.-i. Takada, "Applying the tree algorithm with prioritization for body area networks," in *Proc. of IEEE ISADS*, 2011.