Enhancing 5G Core Network Performance through Optimal Network Fragmentation and Resource Allocation

Madhava Rao Maganti
Department of CSE, College of Engineering and Technology, Acharya Nagarjuna University, India
madhavaraomaganti@gmail.com (corresponding author)

Kurra Rajashekar Rao
Department of CSE, Usha Rama College of Engineering, India
krr_it@yahoo.co.in

Received: 12 March 2024 | Revised: 27 March 2024 | Accepted: 10 April 2024

ABSTRACT

The rise of 5G technology has brought with it a surge in diverse services with demanding and varying requirements. Network fragmentation has emerged as a critical technique to address this challenge, enabling the creation of virtual network segments on a shared infrastructure, allowing for efficient resource utilization and improved performance. This paper investigates the potential of network fragmentation, combined with optimized resource allocation, to enhance the performance of 5G core networks. A novel framework that integrates these two techniques is proposed. The former takes into account factors, such as network traffic patterns, service requirements, and resource availability. The framework aims to optimize network performance metrics, namely throughput, latency, and resource utilization. The experimental results demonstrate the effectiveness of the proposed framework, showcasing a significant improvement in overall network performance, paving thus the way for efficient and robust 5G service delivery.

Keywords-5G RAN; 5G core; 5G network fragmenting; resource allocation; 5G applications

I. INTRODUCTION

The introduction of 5G technology has brought about a significant paradigm shift in wireless communication, promising ultra-high speed, low latency, and massive connectivity. Network fragmentation has emerged as a key technique in 5G networks to accommodate the diverse requirements of various applications and services. Network fragmentation allows the creation of multiple logical networks, or slices, on a shared physical infrastructure. Each slice can be tailored to specific service requirements, ensuring efficient resource utilization and improved network performance. Authors in [1] proposed a framework for optimizing the placement of Network Functions (NFs) in a virtualized cellular core network. The framework aims to minimize the overall network latency and energy consumption while considering various constraints, such as resource availability, NF dependencies, and traffic patterns. Authors in [2] proposed network slicing and fragmentation for creating customized and efficient networks for different applications in 5G and future wireless technologies. A RAN fragment and a core network fragment must be paired in order to produce an end-to-end fragment. There are two possible configurations for the relationship between these fragments: 1-to-1 and 1-to-M. This can really mean that a core fragment can be connected to numerous RAN fragments, and a single RAN fragment can be linked to various core fragments [3, 4]. The goal of network fragmentation is to improve the performance of the network by reducing the amount of traffic that flows through each fragment by assigning resources to the fragments in a way that minimizes the amount of traffic that flows between them. A demonstration of the relationship between the core and RAN fragments along with resource allocation is shown in Figure 1 in which two networks have been fragmented into four fragments, A, B, C, and D. Each fragment has been assigned a resource, which is represented by the color of the fragment. The resources are labeled as 1, 2, 3, and 4. In this example, the resources have been assigned to the fragments in a way that minimizes the amount of traffic that flows between fragments A and B, and between fragments C and D, because fragments A and B are connected by a high-bandwidth link, and fragments C and D are connected by a low-bandwidth link. By fragmenting the network and assigning resources to the fragments in this way, the performance of the network has been improved.

The main focus when it comes to fragment allocation is on intra-fragment isolation, that is, the physical division of Virtual

Private Network Functions (VPNFs) inside a fragment. As hosting the complete fragment on a single server that is compromised or unavailable presents dangers, this division is essential for improving reliability. To mitigate the effects of a partial compromise or unavailability of the network, the fragment operator employs intra-fragment isolation measures, allowing for recovery from such events and finding the best way to offload tasks and allocate resources in 5G edge computing to improve performance [5].

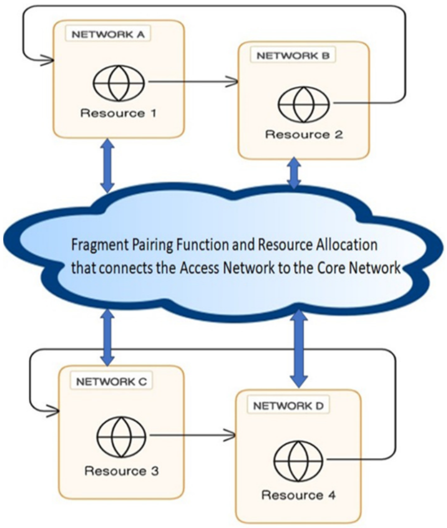
NETWORK A
Resource 1
NETWORK B
Resource 2
Fragment Pairing Function and Resource Allocation that connects the Access Network to the Core Network
NETWORK C
Resource 3
NETWORK D
Resource 4

Fig. 1. Relationship between the core and RAN fragments along with resource allocation.

This study intends to improve reliability and security by enabling on-demand spatial separation between different VPNFs inside a fragment. End-to-end latency considerations are also discussed in the second point. Ensuring compliance is crucial since 5G networks have strict criteria for end-to-end latency. This is especially important when enabling real-time applications like autonomous driving and medical services. The focus is only placed on the end-to-end latency for a core network fragment, but the 5G network is expected to ensure the end-to-end latency for certain applications across the network. This work's primary goal is to optimize fragment allocation in 5G core networks, particularly in the virtual Evolved Core (vEC). This is achieved by applying the VPNF deployment methodology [1]. The contributions of this study address three primary goals: (1) ensuring end-to-end latency, (2) providing intra-fragment isolation for efficient fragment allocation, and (3) identifying the shortest delay path between different parts of a fragment. The central objective is to determine the optimal way to partition a core network fragment in 5G networks. This optimization problem is formulated deploying Mixed-Linear Integer Programming (MILP), considering fundamental requirements which are essential for the distribution of a 5G network fragment. To ensure varying levels of reliability, this study integrates the demand for physical separation among numerous components inside a fragment. Furthermore, the optimization model certifies adherence to end-to-end latency constraints imposed on a fragment of the core network.

The construction of 5G network fragments is dynamic, meaning a fragment can have a modifiable number of components, requiring flexible service chaining. A fragment can include various components, such as User Plane Functions (UPFs) with adaptable service chaining, Security Anchor Function (SEAF), Session Management Function (SMF), Application Function (AF), and Authentication Server Function (AUSF). While the optimization model introduced handles significant aspects of 5G network fragment allocation, it is important to recognize that additional attributes and requirements necessary for a complete end-to-end 5G fragment instantiation fall outside the scope of this paper.

II. RELATED WORK

Network fragmenting aims to allocate physical network resources among multiple networks while minimizing resource usage. Authors in [6] presented a realistic application of network fragmenting with an emphasis on efficiency. Authors in [1] formulated a linear programming framework for VPNF placement within the LTE core network, seeking a compromise between optimality and computational complexity. Authors in [7] proposed an optimization model for VPNF placement in a software-defined networking-based 5G mobile-edge cloud, emphasizing disperse VPNF placement across secure data centers for resilience. Authors in [8] incorporated cost considerations into their allocation algorithm for optimal VPNF placement in mobile virtual cores. Authors in [9] examined different mathematical methods used to optimize the planning and deployment of 5G networks for better performance and efficiency. In [10], a secure and trustworthy system for cloud computing using blockchain technology, which creates a hierarchical structure for managing trust among different users and devices, was proposed. Authors in [11] analyze various data management techniques (clustering, aggregation, compression, encryption, authentication, and gathering) and key management schemes for Wireless Sensor Networks (WSNs). In [12], the genetic algorithm was applied to optimize data operation routes in WSNs, extending network lifetime by conserving energy. In [13], an analytical model compared multipath and single-path CBRP routing protocols in Mobile Ad-hoc Networks (MANETs), demonstrating that multipath routing reduces network congestion and improves end-to-end delay and queue length [13]. Authors in [14] attempted to improve the performance of 4G/5G wireless systems by implementing wavelets for data transmission. Authors in [15] explored the efficient allocation of slices in 5G core networks.

III. MILP FORMULATION

In this section, the optimization model used in this paper, which is based on the framework presented in [1], is analyzed. That study focused on the best location of network services with resource load balancing, mainly addressing the LTE cellular core. The authors relaxed several Mixed-Linear Integer Programming (MILP) restrictions and transformed the optimization problem into a Linear Programming (LP) problem in order to reduce time complexity. To attain the best possible fragment allocation within the 5G core network, the current study expands on their concept. In order to improve dependability, this work aims to strategically distribute VPNFs within the 5G core network fragment while maintaining intra-fragment isolation. The end-to-end core fragment latency, which addresses a basic 5G network requirement, is also guaranteed. In the MILP formulation, the network model and variables described in [1] are employed. Every request in this paradigm has a corresponding computation demand (g^i) and bandwidth need (gij). Specifically customized to the fragment request of this sdtudy, the end-to-end latency (L_{E2E}) and the desired intra-fragment isolation (reliability) between the VPNFs designated as $Qrel$ are taken into account. The objective function employed is:

$$\text{Minimize} \sum_{i \in v_F} \sum_{u \in v_S} \left(1 - \frac{r_u}{r_{v,max}}\right) g^i x^i r_u^i + \sum_{(i,j) \in E_F} \sum_{(u,v) \in E_S} L_{uv} f_{uv}^{ij} \qquad (1)$$

subject to:

$$\sum_{i \in v_F} x_u^i \leq Q_{rel, \forall U} \in v_{s,Q_{rel}} = 1,2,3 \ldots \qquad (2)$$

$$\sum_{(i,j) \in E_F} \sum_{(u,v) \in E_s} \left(\frac{f_{uv}^{ij}}{gij} L_{uv}\right) + \sum_{i \in V_F} a_i \leq L_{E2E} \qquad (3)$$

$$\sum_{i \in V_F} g^i \leq \sum_{u \in V_S} r_u \qquad (4)$$

$$\sum_{(i,j) \in E_F} gij \leq \sum_{(u,v) \epsilon E_S} r_{uv} \qquad (5)$$

Assigning incoming fragment requests to the least used server while reducing the total path delay is the aim of objective function (1). The first term, which focuses on allocating computing demands to the physical servers that are used the least, is comparable to the objective function described in [1]. The argument *ui* is included to prevent the mapping of VPNF/server combinations that are not practical. On the other hand, for virtual linkages $(i,j) \in E_F$, the second term chooses a different path. Here, the physical link delay (L_{uv}), which is dependent on link utilization, is taken into account. L_{uv}, as defined by (6), is computed as the initial delay $L_{uv,init}$, assigned to the link $(u,v) \in E_s$. Finding the path with the least delay between the fragment components and allocating a network fragment to the servers that are not used much is ensured by minimizing both terms in the objective function. Notably, the idea of choosing the least delay path was not included in [1].

$$L_{uv} = \left(1 - \frac{\gamma_{uv}}{\gamma_{uv,max}}\right) 25ms + L_{uv,ini} \forall (u,v) \in E_s \qquad (6)$$

The core and RAN fragments are connected via a transport network. The core fragment is responsible for control plane functions, such as session management, mobility management, and authentication. The RAN fragment is responsible for data plane functions like radio access and packet forwarding. The core and RAN fragments interact with each other through a set of interfaces. These interfaces are defined by the 3GPP standards. The most important interfaces are the S1 interface between the core and RAN fragments and the X2 interface between neighbouring RAN fragments. The Resource allocation problem in a RAN fragment can be formulated as a miMILP problem. The objective of the MILP problem is to minimize the total cost of the network while satisfying a set of constraints. The following are the MILP formulation equations for the resource allocation problem in a RAN fragment:

$$\text{Minimize } f(x) = \sum_{i=1}^{N} C_i X_j \qquad (7)$$

subject to:

$$\sum_{i=1}^{N} X_i \leq K \qquad (8)$$

$$\sum_{i=1}^{N} a_i x_j \leq P \qquad (9)$$

$$\sum_{i=1}^{N} b_i x_i \leq C \qquad (10)$$

$$x_i \geq 0, \forall i \qquad (11)$$

An objective function subject to several MILP restrictions is the subject of the current optimization problem. Constraints (2)–(5) and (9)–(10) from [1] are added to the ones described in this paper. The second constraint deals with the requirement that every VPNF should be assigned to different physical servers in order to guarantee the required degree of reliability, which is represented by $Qrel$, as required by the fragment. Emphasising the necessity of end-to-end latency in the 5G network, constraint (3) is applied to enforce the defined delay requirements for the core network fragment. This restriction takes into account the processing delay of each VPNF, represented by α_i, as well as the delay experienced over the entire path. Constraints (4) and (5) ensure that there is enough processing and bandwidth capacity available throughout the entire data center to support the fragment creation request, preventing partial or incomplete fragment component assignments. This optimization problem aims to find a cost-effective allocation of resource blocks while satisfying the constraints on power consumption, bandwidth consumption, and resource block availability. The decision variables *x*, *i* represent whether each resource block is used or not. The objective is to minimize the total cost of the network while adhering to the aforementioned constraints. This problem is commonly encountered in resource allocation in networking, where the goal is to optimize the utilization of limited resources efficiently.

IV. RESULTS AND DISCUSSION

MATLAB was utilized to simulate the 5G core network and fragment queries in order to verify the optimization model. AMPL was used to model the optimization technique, and CPLEX 12.9.0.0 was the MILP solver that was put into service. The gear employed for the optimization method assessment included an Intel Core i7 processor clocked at 3.2 GHz with 32 GB of RAM. A total of 400 physical terminal servers that could host different kinds of VPNFs were simulated. The evaluation's parameters are listed in Table I. The $Qrel$ parameter was adjusted to change the intra-fragment isolation level during the simulations. The maximum number of VPNFs that can be installed on a single physical server is determined by this option. For a given fragment inside the current network state, the optimization model guarantees compliance with the specified compute resources, bandwidth resources, and end-to-end latency. The remaining bandwidth and processing resources were updated after allocating each fragment. Notably, in network congestion, for instance, the flow link delay L_{uv} may be dynamic and adjust to the existing status of the network. Nevertheless, this work's simulations did not take this dynamic adjustment into account.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Resource capacity/server ($r_{u,max}$)	14.0 GHz
Total terminal servers (v_s)	400
Total fragment requests	400
Q_{rel}	0-10
VPNF/fragment (v_f)	10
Bandwidth request/fragment(g^{ij})	150 Mbps
VPNF resource request/fragment(g^i)	3.2 GHz
α^i	0.7 ms
ε	10^{-10}

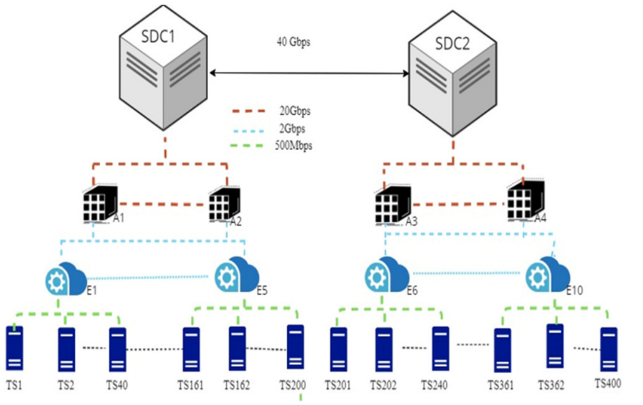

Fig. 2. Network model. TS1-TS400, E1 to E10, A1 to A4, and SDC1-SDC2 represent Console Server, Edge, Amassing and Secure Datacenter switches, respectively.

Two configurations for bandwidth capacity between servers and access switches were used in this study's simulations. The bandwidth capacity is set to 500 Mbps in the first setup, which is portrayed in Figure 2. In this case, the whole performance optimization is limited by the resource capacity that is available (resource bound). Thus, instead of link speed being the limiting element in fragment allocation for the simulated fragment requests, it is the resource capacity of the real servers. The bandwidth capacity is set to 150 Mbps (bandwidth constrained) in the second setting. Here, the accessible data transfer rate capacity between the servers and access switches limits the total system performance. It is significant to note that, unless otherwise indicated, the simulation setup is regarded as "Resource bound."

A. Intra-Fragment Isolation

To reduce the effect of the end-to-end latency (L_{E2E}) on the outcomes, it was set to a relatively high value (500 ms) during the simulation's first phase. The intra-fragment isolation levels (*Qrel*) were then adjusted. The overall average system utilization for accepted requests, bandwidth, and resources at varying levels of intra-fragment isolation is depicted in Figure 3(a). The system has relatively low bandwidth utilization since it is resource-bound and exhibits higher overall system bandwidth than the entire requested bandwidth. The bandwidth utilization rises when fragments request intra-fragment isolation with *Qrel* < 4 because all VPNFs are forced to communicate over physical links.

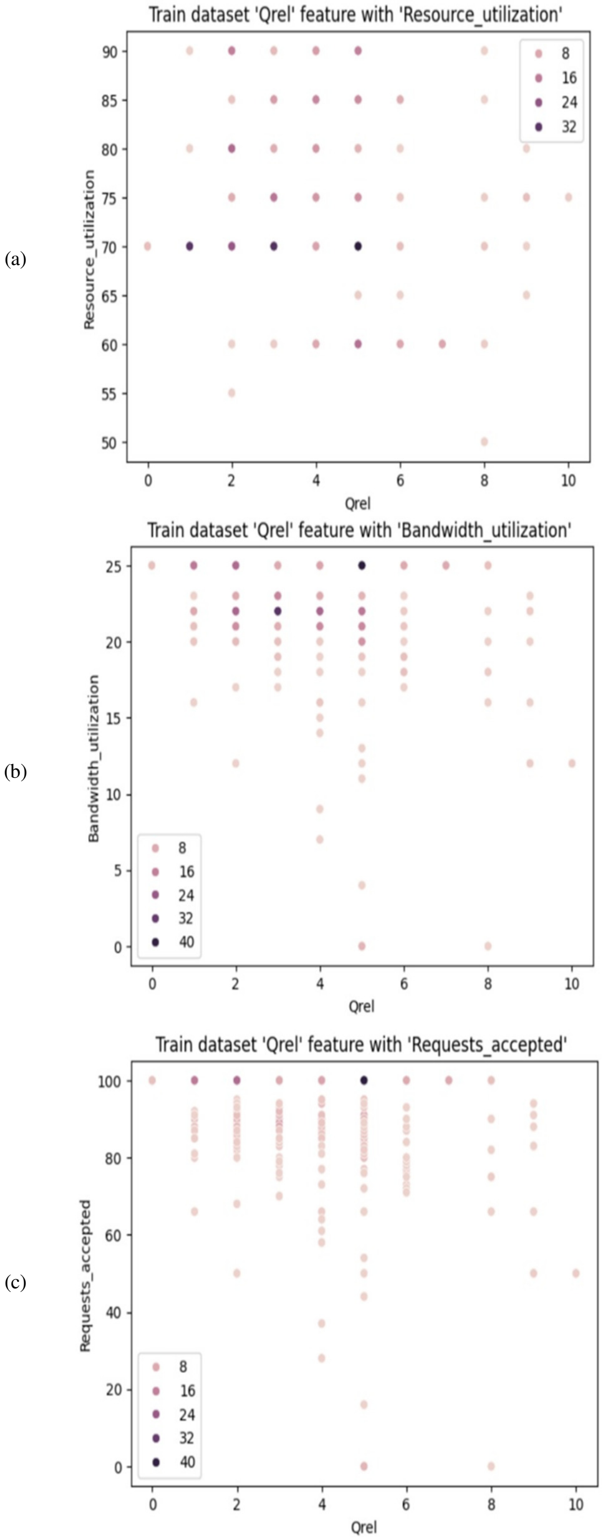

Fig. 3. (a) Resource utilization for varying levels of *Qrel*, (b) bandwidth utilization for varying levels of *Qrel*, (c) requests accepted for varying levels of *Qrel*.

In contrast, reduced network utilization results from reducing the intra-fragment isolation requirement ($Qrel > 4$). This is explained by the fact that more VPNFs can be allocated to the same physical server, which reduces network activity. Additionally, VPNFs can now communicate with one another without the need for physical communication links. Still, there is very little variation in resource usage and approved requests for different $Qrel \geq 2$ levels. This study also contrived a different architecture in which the bandwidth (bandwidth bound) limited the system. The total system utilization for bandwidth is illustrated in Figure 3(b). The bandwidth-bound arrangement results in much worse system performance when compared to the resource bound scenario (Figure 3(a)). Some interesting results about the accepted requests are displayed in Figure 3(c). It is evident that the optimization algorithm requires more time to search an ideal solution for fragment component allocation and to find the optimal path with the least delay when the criteria for intra-fragment isolation increase. These changes can be attributed, in part, to the optimization algorithm's ability to distribute many integrated features of a unified system when the intra isolation criteria are more flexible. This eliminates the need to discover optimal pathways between the components. In Figure 3(c), this behavior is clearly seen when $Qrel > 4$. However, a notable difference in solver performance is noticed when a fragment indicates that no more than double or thrice VPNFs may be installed on a standalone server. These simulations are run several times with different parameter values and almost the same results were attained each time. However, the unusual behavior for $Qrel = 2$ and $Qrel = 3$ is still not understood.

B. End-to-End Latency

Variations end-to-end latency requirements were included in the second portion of the simulation. The fact that simulations were run for $Qrel = 0$ to $Qrel = 10$ should be emphasized, even though the graphs only display the results for a small number of $Qrel$ values. The resource utilization is significantly influenced by the end-to-end latency parameter. This is especially noticeable when setting $Qrel \leq 2$, which results in fewer possible solutions, as manifested in Figure 4.

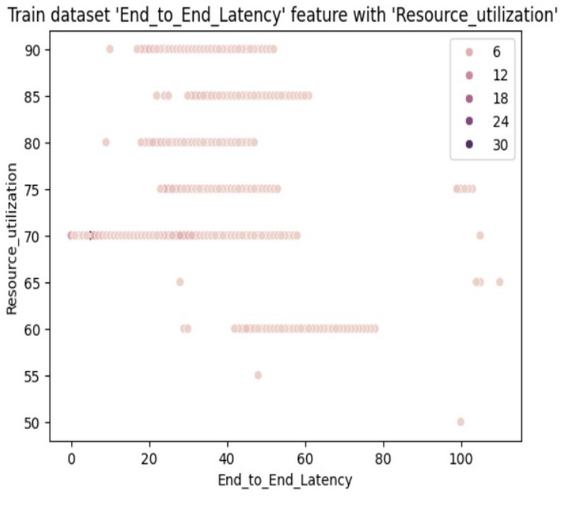

Fig. 4. Resource utilization for variations of L_{E2E} latency requirements.

When the L_{E2E} is adjusted to values equal to or greater than 150, however, this effect is less noticeable. It is noteworthy that the behavior of the resource utilization and acceptance ratio of the requests (which is not shown in this context) are the same. Figure 5 indicates that there is little effect of the variations of L_{E2E} requirements on the total bandwidth utilization for all $Qrel$ levels. Figure 6 discloses the accepted requests for the variations of the L_{E2E} requirements. It is observed that this behavior is constant for all intra-fragment isolation levels.

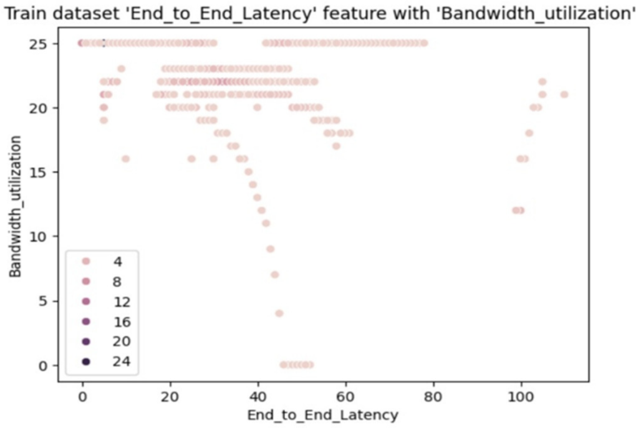

Fig. 5. Variations in L_{E2E} requirements and their impact on bandwidth utilization.

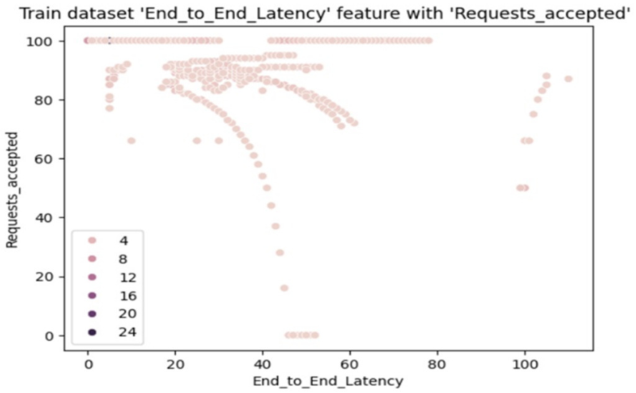

Fig. 6. Variations of L_{E2E} requirements and accepted requests.

V. CONCLUSION

The allocation of network slices and resources is crucial for designing and operating 5G core networks effectively. This paper offers a detailed analysis of the newest advancements and research in this field. Various approaches for network fragmentation, techniques for allocating resources, optimization algorithms, and emerging trends have been covered. This survey equips network operators and researchers with valuable insights to create effective and flexible solutions for 5G core networks' network fragmenting and resource allocation challenges. Consequently, they can enhance the performance and quality of service for a wide range of applications and services. The main focus of this paper was given on recent advances in network slicing techniques, resource allocation methods, optimization algorithms, and emerging trends. By delving into these aspects, the paper provides a comprehensive understanding of the current state-of-the-art methods and identifies potential areas for future research and development.

REFERENCES

[1] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, "Network function placement on virtualized cellular cores," in *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, Bengaluru, India, Jan. 2017, pp. 259–266, https://doi.org/10.1109/COMSNETS.2017.7945385.

[2] M. Taheribakhsh, A. Jafari, M. M. Peiro, and N. Kazemifard, "5G Implementation: Major Issues and Challenges," in *2020 25th International Computer Conference, Computer Society of Iran (CSICC)*, Tehran, Iran, Jan. 2020, pp. 1–5, https://doi.org/10.1109/CSICC49403.2020.9050110.

[3] Q. Li, G. Wu, A. Papathanassiou, and U. Mukherjee, "An end-to-end network slicing framework for 5G wireless communication systems." arXiv, Aug. 01, 2016, https://doi.org/10.48550/arXiv.1608.00572.

[4] *Network Slicing for 5G Networks and Services*. Nov 2016: 5G Americas.

[5] X. Liu and G. Zhang, "Joint Optimization Offloading and Resource Allocation in Vehicular Edge Cloud Computing Networks with Delay Constraints," in *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, Shanghai, China, Sep. 2020, pp. 363–368, https://doi.org/10.1109/PIC50277.2020.9350840.

[6] V. Sciancalepore, L. Zanzi, X. Costa-Pérez, and A. Capone, "ONETS: Online Network Slice Broker From Theory to Practice," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 121–134, Jan. 2022, https://doi.org/10.1109/TWC.2021.3094116.

[7] R. Ford, A. Sridharan, R. Margolies, R. Jana, and S. Rangan, "Provisioning low latency, resilient mobile edge clouds for 5G," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Atlanta, GA, USA, Feb. 2017, pp. 169–174, https://doi.org/10.1109/INFCOMW.2017.8116371.

[8] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," in *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, London, UK, Apr. 2015, pp. 1–9, https://doi.org/10.1109/NETSOFT.2015.7116162.

[9] M. M. Ahamed and S. Faruque, "5G Network Coverage Planning and Analysis of the Deployment Challenges," *Sensors*, vol. 21, no. 19, Jan. 2021, Art. no. 6608, https://doi.org/10.3390/s21196608.

[10] H. Yang, J. Yuan, H. Yao, Q. Yao, A. Yu, and J. Zhang, "Blockchain-Based Hierarchical Trust Networking for JointCloud," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1667–1677, Mar. 2020, https://doi.org/10.1109/JIOT.2019.2961187.

[11] M. B. Apsara, P. Dayananda, and C. N. Sowmyarani, "A Review on Secure Group Key Management Schemes for Data Gathering in Wireless Sensor Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 1, pp. 5108–5112, Feb. 2020, https://doi.org/10.48084/etasr.3213.

[12] A. Rajab, "Genetic Algorithm-Based Multi-Hop Routing to Improve the Lifetime of Wireless Sensor Networks," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7770–7775, Dec. 2021, https://doi.org/10.48084/etasr.4484.

[13] M. A. Mahdi, T. C. Wan, A. Mahdi, M. a. G. Hazber, and B. A. Mohammed, "A Multipath Cluster-Based Routing Protocol For Mobile Ad Hoc Networks," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7635–7640, Oct. 2021, https://doi.org/10.48084/etasr.4259.

[14] B. Domathoti, C. Ch, S. Madala, A. A. Berhanu, and Y. N. Rao, "Simulation Analysis of 4G/5G OFDM Systems by Optimal Wavelets with BPSK Modulator," *Journal of Sensors*, vol. 2022, Sep. 2022, Art. no. e8070428, https://doi.org/10.1155/2022/8070428.

[15] D. Sattar and A. Matrawy, "Optimal Slice Allocation in 5G Core Networks," *IEEE Networking Letters*, vol. 1, no. 2, pp. 48–51, Jun. 2019, https://doi.org/10.1109/LNET.2019.2908351.