LLM-Assisted Constraint-Aware Pipeline (CAP) for Scalable and Optimized SFC Placement in Multi-Generation Networks

Author 1, Author 2, Author 3, Author 4
Department of Computer Science, University Name, Country
Email: author1@example.com, author2@example.com, author3@example.com, author4@example.com

I. Related Work

Service Function Chaining (SFC) placement has been a critical research area in the domain of next-generation networks, particularly in the context of 5G and beyond. Several studies have explored different aspects of SFC deployment, including computational efficiency, latency minimization, resource allocation, and service continuity. However, these approaches often face limitations such as high computational overhead, lack of adaptability to real-time network changes, and inadequate predictive mechanisms for handling dynamic traffic fluctuations.

One of the early works on distributed SFC placement is presented in [1], where the authors proposed a heuristic-based Distributed Service Function Chaining (DSFC) framework utilizing the Kariz algorithm. Their approach focused on reducing the dependency on centralized controllers by distributing SFC placement decisions across the network. While the method effectively improved computational efficiency and distributed decision-making, it lacked real-time adaptability and did not incorporate traffic forecasting, making it less suitable for highly dynamic environments where network conditions change rapidly.

Another significant study in the area of data center network optimization is found in [2], where an Analytic Hierarchy Process (AHP)-based model was employed to design efficient network topologies. This approach evaluated various data center topologies based on multiple criteria, such as cost, reliability, and network performance. While the AHP-based decision-making model provided an effective way to balance multiple factors in static scenarios, it did not support real-time dynamic SFC deployment, which is essential for evolving network conditions in multi-generation environments.

Bandwidth management plays a crucial role in ensuring efficient network operation, especially in the presence of long-duration large flows. In [3], an adaptive bandwidth control mechanism was proposed to handle such flows without negatively impacting short-duration latency-sensitive traffic. The approach dynamically adjusted bandwidth allocation based on observed traffic patterns, ensuring fair resource distribution among competing flows. However, the method was primarily focused on bandwidth optimization at an individual link level rather than a comprehensive SFC placement strategy that considers multiple interconnected services.

Energy efficiency has also been a key concern in SFC placement, especially in large-scale multi-domain networks. In [4], a reinforcement learning-based framework called SCHE2MA was introduced, which employed a multi-agent auction mechanism for energy-aware SFC orchestration. The model aimed to reduce the power consumption of Virtual Network Functions (VNFs) while maintaining service performance. Despite the significant energy savings achieved, the framework required extensive training for deep reinforcement learning models, making it computationally expensive and unsuitable for real-time adaptive deployment.

Another work that explored end-to-end slice orchestration in 5G networks is presented in [5], where the authors formulated multiple Mixed Integer Linear Programming (MILP) models and heuristic approaches for optimal VNF placement. The study primarily focused on minimizing resource costs, bandwidth consumption, and service migration overheads. While MILP-based optimization provided high accuracy in static environments, its high computational complexity limited its application in large-scale, real-time networks where rapid decision-making is necessary.

Latency-sensitive SFC placement has been extensively researched to ensure quality-of-service (QoS) compliance in delay-sensitive applications. In [6], an Integer Linear Programming (ILP)-based strategy, combined with a heuristic approach, was introduced to optimize latency and resource utilization by efficiently placing VNFs along network paths. While this method minimized end-to-end latency, its high computational complexity posed scalability challenges, necessitating heuristic solutions to improve feasibility for large-scale implementations.

A step further was taken in [7], where the authors incorporated user mobility into the SFC placement model using a Mixed Integer Linear Programming (MILP) approach combined with a heuristic method. Their goal was to minimize service disruptions caused by frequent handovers while ensuring efficient resource allocation. The study successfully reduced service degradation in mobile networks; however, its heuristic-based methodology still exhibited limited adaptability in highly dynamic environments, as it did not proactively predict workload fluctuations.

A. Our Contribution

Our proposed LLM-assisted Constraint-Aware Pipeline (CAP) addresses these challenges by integrating a constraint-aware path selection mechanism with a resource-aware microservice allocator, enabling real-time, adaptive, and automated SFC deployment. Additionally, we incorporate SARIMA-based traffic forecasting to predict workload variations, ensuring proactive resource allocation. By leveraging LLMs for automated configuration generation, our approach reduces deployment overhead while optimizing microservice placement across multi-generation networks. This comprehensive framework offers a scalable, intelligent, and future-proof solution for efficient SFC placement in dynamic network environments.

II. References

References

[1] M. Ghaznavi, N. Shahriar, S. Kamali, R. Ahmed, and R. Boutaba, "Distributed Service Function Chaining," IEEE Journal on Selected Areas in Communications, vol. 35, no. 11, pp. 2479–2492, Nov. 2017.

[2] N. Kamiyama, "Designing Data Center Networks using Analytic Hierarchy Process," IEEE ICCCN, 2010.

[3] R. Kawahara, T. Mori, N. Kamiyama, S. Harada, and H. Hasegawa, "Adaptive Bandwidth Control to Handle Long-Duration Large Flows," IEEE ICC, 2009.

[4] A. Dalgkitsis et al., "SCHE2MA: Scalable, Energy-Aware, Multi-Domain Orchestration for Beyond-5G URLLC Services," IEEE Journal, 2022.

[5] D. Harutyunyan et al., "Orchestrating End-to-End Slices in 5G Networks," IEEE CNSM, 2019.

[6] Author(s), "Latency-Aware Service Function Chain Placement in 5G Mobile Networks Using ILP," Conference/Journal, 2019.

[7] Author(s), "Latency and Mobility–Aware Service Function Chain Placement in 5G Networks Using MILP and Heuristic Approach," Conference/Journal, 2020.