Optimizing 5G Network Slices: LSTM and Game Theory Synergy

Emmanuel J Samson†, Kamrul Hasan†, Liang Hong†, Imtiaz Ahmed††, Henry Onyeka†, Sachin Shetty†††
† Tennessee State University, Nashville, TN, USA
†† Howard University, Washington, DC, USA
††† Old Dominion University, Norfolk, VA, USA
Email: {*esamson, mhasan1, lhong, honyeka*}@tnstate.edu, sshetty@odu.edu, imtiaz.ahmed@howard.edu

***Abstract*—Advancements in fifth generation (5G) technology have seen a rise in adapting network slicing (NS) to differentiate the services offered in different network segments due to differences in performance requirements and traffic characteristics. With the growing demand for network automation, several works have explored the concepts of intelligent network slicing. This work, however, demonstrates an approach to traffic prediction that informs network function (NF) resource demands based on the traffic pattern studied in the network slices (NSs). Using traffic volume as the basis for sharing resources among NFs in the NSs, Long Short-Term Memory (LSTM) algorithm is used to study and predict the traffic patterns. This prediction is then used as input to the resources allocation algorithm, that is based on game theory, where the resources are allocated dynamically and fairly amongst the NSs. The work demonstrates a foundation on which all other resources allocation–in the Radio Access Network (RAN), Transport Network (TN), and Core Network (CN)–can be based on to formulate strategic resource sharing amongst NFs.**

***Index Terms*—5G, NF Placement, Intelligent Network Slices**

I. Introduction

The development and advancement of fifth generation (5G) technology has allowed it to incorporate Network Slicing (NS) in its operations. The differentiation in the kind of services offered by the wireless networks has incorporated multi-modal services, especially those based on Internet services.

To offer services, cellular communication networks have three major technical areas to consider: radio access network (RAN), transport network (TN), and core network (CN). The functions of these technical areas are the ones that make telecommunication services and network access possible.

While these areas work traditionally, the current advancements in technologies and wireless communication networks and incorporating NS in 5G requires a new way of looking at network operations. The 5G service-based architecture (SBA) promises that the network should be deployed depending on the service requirements [1]. In general, expertise in the area would suffice in coming up with all the service level agreements (SLAs), configurations, and necessary key performance indicators (KPIs) that would satisfy the new requirements [2]. However, the physical infrastructure to support various networks is always limited, hence the need for the network slices (NSs) to share the infrastructure.

To make an efficient use of the system bandwidth, for example, traditional ways of creating virtual local area networks (VLANs) or using multi-protocol label switching (IP/MPLS) approaches would typically suffice. However, the changes in traffic metrics limit their usefulness for operations in varying network characteristics [3]. Therefore, different efforts have been made to incorporate intelligence in network operations concerning NS in the RAN, TN, and CN [4]–[6]. However, there has not been comprehensive NS optimization solutions that comprehensively address the heterogeneity and dynamicity in the NSs and looks to expand the shared physical and virtual resources.

NSs provide their functions by defining service chains (SC) that incorporate network functions (NFs). With 5G and beyond technologies, NFs run on shared servers, decoupled from specialized hardware, using network function virtualization (NFV), virtual network functions (VNFs), and cloud network functions (CNFs). This paper focuses on optimizing network slices by managing the shared memory and bandwidth of clustered servers. It illustrates how analyzing network traffic can help understand resource demands and allocate them fairly based on the characteristics of services within an NS.

This paper starts by reviewing the literature on network slicing in Section II, exploring the concepts of network slicing in all domains, and concentrating on intelligent slicing in the succeeding subsection. Section III presents the methodology used to set up the study. The results are discussed in Section IV. We conclude the paper in Section V.

II. Related Works

A. Intelligent Network Slicing

Continuous improvements in fronthaul and backhaul technologies have led to an increase in network users, necessitating constant advancements in provisioning and scheduling algorithms for smooth network operations. Addressing these issues requires considering traffic characteristics, but the heterogeneous and stochastic nature of cellular network traffic complicates ensuring dynamicity and fairness in resource sharing. Network slicing faces similar challenges. Consequently, numerous studies have been conducted on slicing in the three domains of RAN (Radio Access Network), CN (Core Network), and TN (Transport Network).

1) Intelligent RAN and CN Slicing: The RAN is concerned with radio resources sharing, whereas the TN and CN are concerned with transport resources sharing and getting the NFs

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

near the user respectively. To model the sharing and scaling network resources [7], one must consider different approaches in all those domains.

Therefore, from the RAN end, several problems in [8] were identified to validate intelligent approaches in scaling resources in real-time updates as the network operates. First, it identifies and defines the periodic traffic prediction for slices problem. Second, it defines the online resource scheduling problem.

Understanding the resource scheduling prediction requires periodic checks and continuous monitoring of traffic status in the network. In the first problem, [8] defines the periodic traffic prediction by defining the decision period which informs the algorithm that is responsible for prediction of resources that are required for each slice when to run the predictions. The mathematical objective in this problem was to minimize the mean squared error (MSE) that resulted from the actual resource against the actual demands of individual network slices in the network. The allocations are subject to the available boundaries defined by the availability of resources.

Within the 5G core network (5GC), however, three main challenges about orchestration are discussed in [9]: (1) NF placement, (2) scaling strategy, and (3) scaling sequence. However, resources allocation problem is subject to NF placement and resource scaling strategy. The ultimate scaling issue in this case is how much resources should be provisioned for a NF running on a server informed by the traffic volume in it.

The NFs running in the core are separated in their general functions. In this regard, we must point out that we have both signaling NFs, which are concerned with relaying control messages in a 5GC, and control NFs, which are paramount to maintaining the network state and central monitoring of network operations. [10]–[12] explore the access and mobility management function (AMF) scaling issues, and [13]–[16] explore the user plane function (UPF) scaling issues. In their totality, they explore traditional and machine learning/artificial intelligence (ML/AI)-driven techniques to scale these NFs. Based on these findings, we conclude that it is necessary to incorporate NFs scaling issues related to traffic differences within the NSs in the 5G SBA.

Therefore, with changing traffic volume, the allocated resources for the NFs can be optimized to meet traffic demands.

2) Intelligent NF Placement: The decoupling of NFs from specialized hardware the user plane (UP) from the control plane (CP) in 4G and 5G with its accompanying SBA introduced the necessity to rethink how NF resource allocation and placement are performed. Traditionally, this would require the understanding of the traffic patterns [17], hence provisioning resources with respect to busy hour traffic patterns. To use the traffic flow data to inform the network controllers how much resources are required by the NFs and when in the NSs, one must consider the end-to-end (E2E) network operations. To realize the E2E network slicing, therefore, one must define different problems in each domain, hence the effect on NFs placement. In their entirety, though, the issues should address the change in traffic patterns and define a foundation on which those different problems would base. Defining this foundation is the focus of demonstration in this work.

Among many works of interest, [18] stands out as a work that captures the intrinsic nature of traffic flow within the network. While their work concentrates on Open RAN (O-RAN), their idea that captures the hierarchical data centers within the cellular communication network plays a key role in demystifying the NF placement issues within the network when we look at the E2E performance. For this case, utilizing the aggregate traffic in the NSs, and assuming they serve the same network domain, and share the same physical and network resources, we demonstrate how learning the traffic patterns can be utilized to allocate resources for the NFs to ensure dynamic, timely, and fair utilization of resources available.

III. Methodology

The study herein goes through three main steps: traffic data simulation, training traffic prediction model, and network resources allocation. The traffic simulation assumes the aggregate traffic in the NSs, and that while the traffic in the NSs will have different characteristics, the generation assumes same packet characteristics of this data. The steps are discussed in the following subsections.

A. Traffic Data Simulation

This work utilizes the simulation of cellular communication traffic data to generate the data necessary for the study. The data is generated as a time series capturing the gradual increase and decrease of traffic in three different slices modeling the enhaced mobile broadband (eMBB), massive machine-type communication (mMTC) and ultra reliable low latency communication (URLLC) characteristics of the network slices in 5G networks [19]. The parameters in Table I were used.

TABLE I: Traffic Data Simulation Parameters

Parameter	***Value***
Base Traffic for off-peak hours ($\mathcal{T}$)	1000 packets
Peak Hour (t_p)	18 (6 PM)
Iterations	24*30 (24 hrs; 30 days)
Peak Factor (ρ)	2.5
S.D. for random variation (σ)	0.2
Maximum allowable variation in traffic ($\hat{\sigma}$)	0.2

The traffic data generation is a three step process that involves the process in (1) where ω is calculated with an intent to model the transitions of traffic changes. Here, μ_τ is the mean traffic as modeled with the sine equation and η_τ denotes the total traffic that is generated instantaneously. The traffic is modeled as a Gaussian process ensuring that the traffic generated is always greater than 0. Therefore, any traffic generation that is less than 0 is clipped to a 0 in (1).

$$\begin{cases} \omega = \sin(\dfrac{\pi \times (t - t_p)}{12}) + 1, \\ \mu_\tau = \mathcal{T} \times (1 + (\rho - 1) \times \dfrac{\omega}{2}), and \\ \eta_\tau = \mathcal{N}(\mu_\tau, \sigma * \mathcal{T}). \end{cases} \tag{1}$$

Thereafter, this traffic is distributed amongst slices with consideration on the previous traffic volume. In this case, the maximum change allowed in traffic is first computed by obtaining Δ_τ, which is thereafter used to randomly generate a change factor Δ for that particular instance of time. Δ is then used to compute the traffic in a slice at an instance, $\eta^m(t)$, where $m \in \mathcal{M}$ and $\mathcal{M}$ is the total number of slices. Each slice is, therefore, provided with its own traffic volume over the course of traffic data generation.

$$\begin{cases} \Delta_\tau = \eta_\tau * \hat{\sigma} \\ \Delta = [-\Delta_\tau, \Delta_\tau] \\ \eta^m(t) = \eta^m(t-1) + \Delta \\ where 0 < \eta^m(t-1) + \Delta < \eta_\tau \end{cases} \tag{2}$$

This approach generated traffic distribution that is close to the actual telecommunication traffic, which models the gradual changes in traffic, busy hour and different traffic volumes in the eMBB, mMTC, and URLLC slices at certain times of the day, thus resulting to Fig. 1. It is worth mentioning that our simulated traffic assumes that the peak hour is at 6 PM.

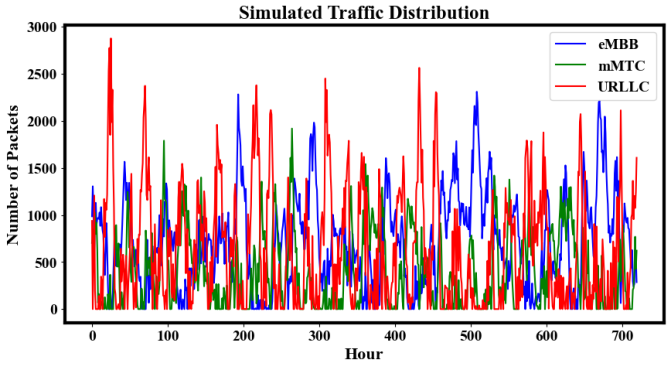


Fig. 1: Simulated Traffic Distribution

B. Training Traffic Prediction Model

Despite the random arrival process of the individual user equipment (UEs) accessing the cellular communication networks, the aggregate traffic in these networks–which is the basis for network function placement–follows the time series trends, where the increase and decrease in traffic can be observed over long periods of time. This is the case for the individual network slices too. Hence, after generating the data and confirming that its shape conforms to the time series trends of the telecommunication data from the literature, the traffic prediction model was devised.

Studying time series data requires that we have longevity in retaining the data in both short and long terms. Long Short-Term memory (LSTM) neural network architecture provides this capability and has been utilized in a wide range of applications [20]. LSTM neural networks are a type of recurrent neural network (RNN) designed to handle long-term dependencies in time series data. Each LSTM cell contains three main gates: the input gate, which controls the extent to which new information flows into the cell state; the forget gate, which determines how much of the past information should be discarded; and the output gate, which regulates the information sent to the next hidden state. With two hidden layers, the network can learn complex patterns by processing sequences through multiple stages of these gates, retaining essential information over long periods and thus improving the accuracy of time series prediction tasks.

Therefore, using Tensorflow, an open source ML framework created by google, the traffic prediction architecture is developed with two hidden layers of 50 neurons each and drop out rate of 20%. Other parameters are tabulated in Table II. This architecture is then replicated to produce 3 separate models subjected to 3 generated datasets, for eMBB, mMTC, URLLC respectively, in order to treat the traffic trends within the network slices differently. The training and validation losses are presented in Fig. 2, where the model is observed resulting into very small errors and has been kept on updating itself over the course of 50 epochs. Consequently, using the considered models result to the mean squared error (MSE) of $0.0082, 0.0097$, and 0.0178 for the eMBB, mMTC and URLLC models respectively.

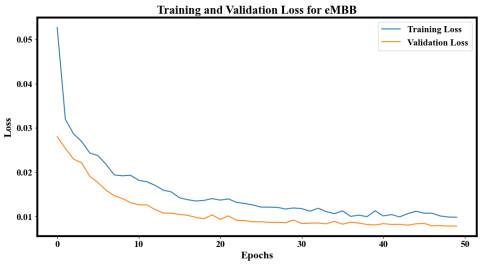
TABLE II: LSTM Model characteristics

Parameter	***Value***
Training set	80%
Validation Set	20%
Batch size	32
Sequence length	24 (hours)
Epochs	50
Hidden Layers	2
Number of Neurons on each layer hidden	50
Dropout rate	20%
Optimizer	Adam
Scaler	MinMaxScaler

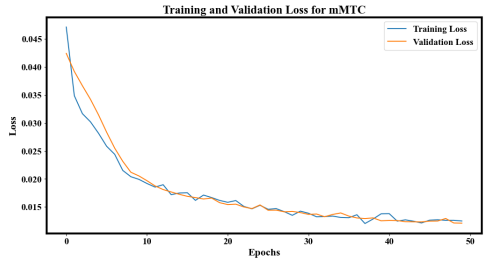
C. Network Resources Allocation

The placement of an NF requires knowledge of its optimal location and resource requirements, such as CPU, memory, and bandwidth. Servers must have sufficient physical resources to deploy and run these NFs, ensuring enough capacity to handle all active functions. NFs compete for these server resources and are organized into service chains (SCs), which outline the sequence of NFs a service must pass through. These service chains define network services (NSs), which vary based on traffic type, subscriber capacity, and service offerings. Consequently, NSs aggregate traffic patterns, which in turn shape the traffic patterns within NFs, guiding the optimization strategies for individual NFs.

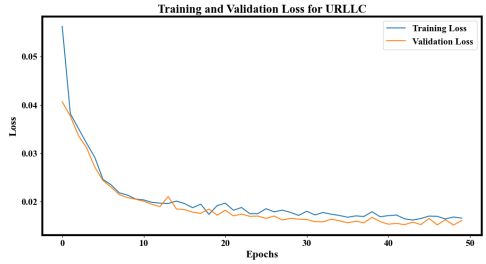
While all the Internet backbone runs on Internet protocol (IP), the quality of service (QoS) levels of these NSs are defined and limited by the reliability, dependability, and bandwidth constraints. Hence, the utility function $U_i(b_i, m_i)$, where b_i and m_i are the bandwidth and memory required for the VNF in slice i, respectively. In (3), α_i and β_i are scaling factors for the bandwidth and the memory, respectively. Furthermore, we note in (4) that the algorithm maximizes the utilization of the resources, and that the utilization of resources is the sum of all individual consumption (of the NFs in the NSs) of the



(a) eMBB Slice

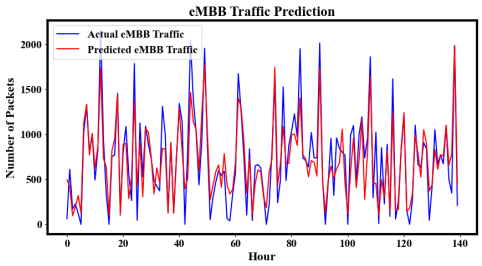


(b) mMTC Slice

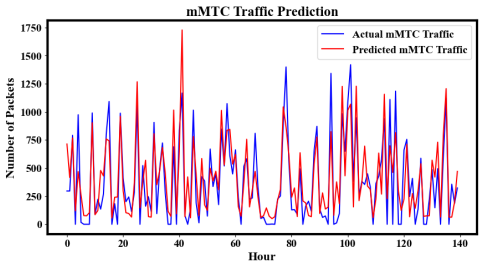


(c) URLLC Slice

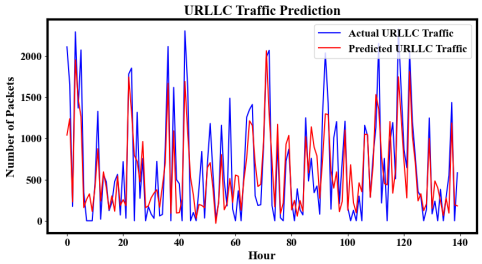
Fig. 2: Training and Validation Losses



(a) eMBB Slice



(b) mMTC Slice



(c) URLLC Slice

Fig. 3: Traffic Predictions in the Slices

available resources while keeping in mind that each individual utilization of the resources is below the maximum configured capacities of the hosting infrastructure as in (5).

$$U_i(b_i, m_i) = \alpha_i b_i + \beta_i m_i \tag{3}$$

$$\max_{\{b_i, m_i\}} \sum_i U_i(b_i, m_i) \tag{4}$$

$$\begin{cases} \sum_i b_i \leq B; \\ \sum_i m_i \leq M; \\ b_i, m_i \geq 0 \end{cases} \tag{5}$$

Game theory approach—Algorithm 1—is utilized, where resource allocation involves strategic decision-making where multiple agents (players) compete for limited resources. It defines the strategic considerations made in order to let the NSs compete for the resources. This follows as game theory has been utilized to find stable states where no agent can improve their outcome by unilaterally changing their strategy. In this work, the players are the NSs whose traffic patterns have been studied and predictions for the next round of traffic volume has been determined. Each NS's objective is to maximize its own payoff, which depends on its own actions and the actions of other NSs. The inputs to the algorithm, $[\hat{\tau}_{NS}]$, m, s, are an array of traffic volume predictions, number of slices, and strategy of resources allocation respectively. On each iteration $t \in T$, the resources allocated for each slice $i \in m$, $r_i^{mem}(t)$ and $r_i^{bw}(t)$, are changed only when the succeeding allocation has at least 10% [1] difference with the preceding one.

It follows that combining game theory with LSTM enhances decision-making in dynamic competitive environments–traffic differences and limited resources in this context–by leveraging the strength of LSTM in predicting future network traffic and game theory in determining stable states for resources allocation. In this context, however, game theory helped to design a fair and efficient foundation for distribution of the resources amongst NFs running in the slices.

IV. Results and Discussion

This work demonstrates the allocation/distribution of two resources available for the NSs. The first is the bandwidth, which is necessary for the transmission of packets to and from the NFs running in the NSs, and the second is the memory, which is at the core of information processing of

[1] 10% was chosen for demonstration purposes. This number can be changed depending on operational constraints.

Algorithm 1 Game Theory Based Resources Allocations

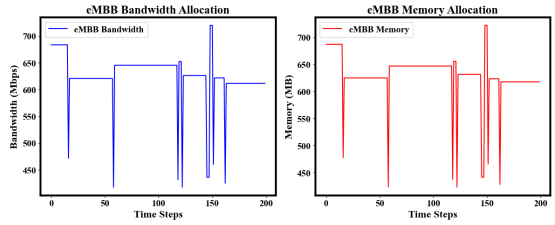
Input: $[\hat{\tau}_{NS}]$, m, s
Output: [r] {r: resources allocation}

set $T \leftarrow length(\hat{\tau}_{NS})$ {Size of an array of predictions}
for i **from** 1 **to** m **do**
  set $r_{i0}^{mem} \leftarrow R/m$ {Memory Allocations}
  set $r_{i0}^{bw} \leftarrow R/m$ {Bandwidth Allocations}
end for
for t **from** 1 **to** T **do**
  for i **from** 1 **to** m **do**
    $\tau_{NS}^{i} \leftarrow \hat{\tau}_{NS}^{i}(t)$ {Predicted traffic at t+1}
    $r_i^{mem}(t) \leftarrow [s \times \tau_{NS}^{i}]$ {Memory allocation strategy}
    $r_i^{bw}(t) \leftarrow [s \times \tau_{NS}^{i}]$ {BW allocation strategy}
    $r(t) \leftarrow [r_i^{mem}(t), r_i^{bw}(t)]$
    if $(|r(t) - r(t-1)| > 0.1 \times r(t-1))$ **then**
      $r \leftarrow r(t)$
    end if
  end for
end for
return r

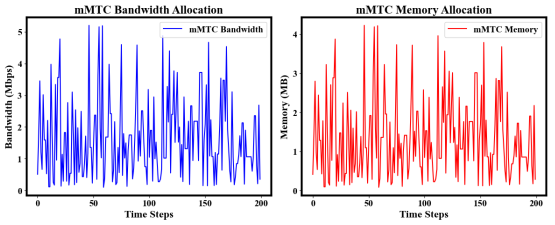
the NFs. In this regard, it is necessary to note that the NFs run in the servers. And, while the NFs capacities are always defined at their deployment, and the requirements are pre-configured, it is in the 5G and beyond vision that the NFs should utilize only the resources they require. For this reason, the presence of a single cluster of servers that hosts the NFs with imaginary $1GB$ of memory and a network interface configured with $1Gbps$ of bandwidth is assumed. With these values, therefore, the resources are allocated by subjecting the predicted traffic in individual slices to Algorithm 1 which resulted into allocations in Fig. 4 and Fig. 5. Considering the operations in the real world, the allocations start from the server level, or the orchestrator in virtualized services, sense before it is extended to the SCs and NSs level, making our approach customizable in all levels to meet required service levels.

From Fig. 4 and Fig. 5, one should note that there are differences in allocations. It is notable from the figures that the eMBB slice in Fig. 4 was allocated with a lot of resources compared to the other two slices. However, in Fig. 5 this changes. Despite the eMBB acquiring more resources than the other slices–individually–the slices are able to acquire reasonable resources. This is due to changing the strategy used to allocating the resources in for the bandwidth and the memory available. In Fig. 4, the strategy (requisition factors), $s_{BW} = \{2, 1, 0.5\}$ and $s_{MEM} = \{1, 1, 0.5\}$ were used for eMBB, mMTC, and URLLC slices respectively. On the other hand, Fig. 5 used $s_{BW} = \{0.4, 0.5, 0.2\}$ and $s_{MEM} = \{0.2, 1, 0.5\}$ for eMBB, mMTC, and URLLC slices respectively. It is, therefore, noted that the resources allocation strategy should be clearly defined to attain reasonable allocations of resources. For this case, the utilization of the slices should inform how this strategy is optimized.

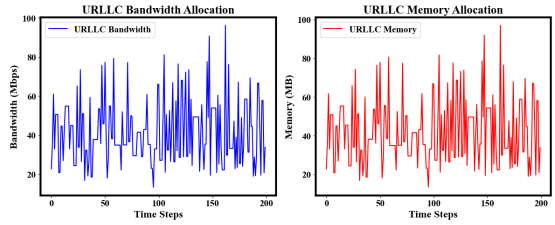
Furthermore, the update cycles in both strategies were found to be reasonable in order to avoid the configuration overheads. It is necessary to note that the time steps that are used here are in hourly basis, and that the resources changes are accepted only when the resulting allocation–that is determined by the projected demand–exceeds 10% of the previous allocation. The effect of this feature is clearly observed from the bandwidth allocations in the eMBB and URLLC slices in Fig. 5.



(a) eMBB Slice



(b) mMTC Slice



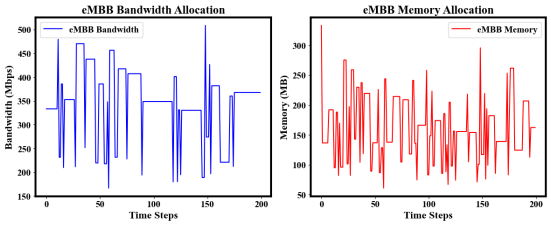
(c) URLLC Slice

Fig. 4: Traffic Prediction Informed Resources Allocation in the Slices

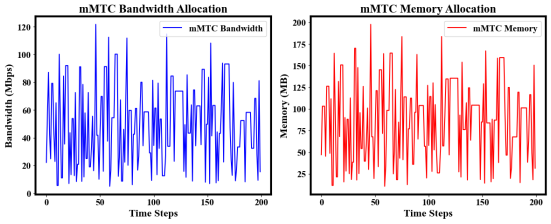
Finally, it is worth noting that, while a very recent work in [21] utilized LSTM in predicting the throughput of several common service models in 5G, it did not concentrate on allocating the resources for the services. Our approach demonstrates this clearly. Clearly, the paper demonstrates a different approach from [22]—which concentrates on resource allocation in the RAN using deep reinforcement learning—demonstrating that resource allocation should happen in real time using the historical traffic data, and attaining better results in less iterations (50 vs 1000) thus reducing the computational time and computational resources load. It is also considered in the algorithm that small differences in traffic would not result in a change in allocated resources.

V. Conclusions and Future Work

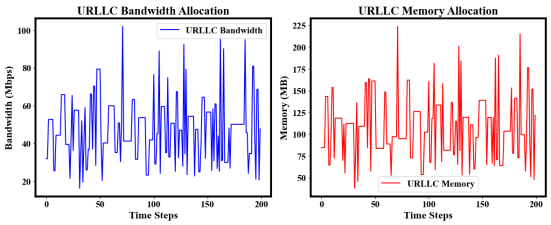
The work herein aimed at optimizing NSs resources allocations. The optimization of the resources is a required feature for 5G and beyond networks, especially considering the ever growing demand in terms of traffic within the networks. With growing automation approaches, this work has explored the



(a) eMBB Slice



(b) mMTC Slice



(c) URLLC Slice

Fig. 5: Traffic Prediction Informed Resources Allocation in the Slices With an Updated Strategy

synergy between LSTM and game theory approach in allocating resources for the NFs serving the NSs. It has demonstrated that traffic prediction can be done with high accuracy and the resources allocations only change when it is necessary. Compared to recent works that have tackled the problem using deep learning, it was observed that the algorithm resulted in better results for predicting traffic with less iterations than the other approaches, and allocating resources better with those predictions.

Therefore, this work can be extended to different domains spanning the RAN, TN, and CN of the 5G and beyond technologies to provision for resources required to run network functions. It is necessary to note that this work assumes that the NFs can run on any server as efforts are still being made to make the NFs cloud-native and hardware independent, which is one of the key technological advancements brought by 5G.

Acknowledgment

This material is based on work supported by the National Science Foundation Award Numbers 2205773, 2219657, and 2219658.

References

[1] J. B. Moreira, H. Mamede, V. Pereira, and B. Sousa, "Next generation of microservices for the 5g service-based architecture," *International Journal of Network Management*, vol. 30, no. 6, p. e2132, 2020, e2132 nem.2132. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nem.2132

[2] M. R. Sama, X. An, Q. Wei, and S. Beker, "Reshaping the mobile core network via function decomposition and network slicing for the 5g era," in *2016 IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–7.

[3] Makhija, Deven, "5g communication systems: Network slicing and virtual private network architecture," *ITM Web Conf.*, vol. 54, 2023. [Online]. Available: https://doi.org/10.1051/itmconf/20235402001

[4] A. Arnaz, J. Lipman, M. Abolhasan, and M. Hiltunen, "Toward integrating intelligence and programmability in open radio access networks: A comprehensive survey," *IEEE Access*, vol. 10, pp. 67 747–67 770, 2022.

[5] Y. Fu, S. Wang, C.-X. Wang, X. Hong, and S. McLaughlin, "Artificial intelligence to manage network traffic of 5g wireless networks," *IEEE Network*, vol. 32, no. 6, pp. 58–64, 2018.

[6] A. A. Gebremariam, M. Usman, and M. Qaraqe, "Applications of artificial intelligence and machine learning in the area of sdn and nfv: A survey," in *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, 2019, pp. 545–549.

[7] C. H. T. Arteaga, A. Ordoñez, and O. M. C. Rendon, "Scalability and performance analysis in 5g core network slicing," *IEEE Access*, vol. 8, pp. 142 086–142 100, 2020.

[8] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, "Intelligent resource scheduling for 5g radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691–7703, 2019.

[9] A. Sheoran, S. Fahmy, L. Cao, and P. Sharma, "Ai-driven provisioning in the 5g core," *IEEE Internet Computing*, vol. 25, no. 2, pp. 18–25, 2021.

[10] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the scalability of 5g core network: The amf case," in *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 2018, pp. 1–6.

[11] A. Chouman, D. M. Manias, and A. Shami, "A reliable amf scaling and load balancing framework for 5g core networks," in *2023 International Wireless Communications and Mobile Computing (IWCMC)*, 2023, pp. 252–257.

[12] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, C. Viho, and D. Darche, "Smart scaling of the 5g core network: An rnn-based approach," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[13] C. Rotter and T. Van Do, "A queueing model for threshold-based scaling of upf instances in 5g core," *IEEE Access*, vol. 9, pp. 81 443–81 453, 2021.

[14] H. T. Nguyen, T. Van Do, and C. Rotter, "Scaling upf instances in 5g/6g core with deep reinforcement learning," *IEEE Access*, vol. 9, pp. 165 892–165 906, 2021.

[15] D. Kumar, S. Chakrabarti, A. S. Rajan, and J. Huang, "Scaling telecom core network functions in public cloud infrastructure," in *2020 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 2020, pp. 9–16.

[16] J. Costa-Requena, A. Poutanen, S. Vural, G. Kamel, C. Clark, and S. K. Roy, "Sdn-based upf for mobile backhaul network slicing," in *2018 European Conference on Networks and Communications (EuCNC)*, 2018, pp. 48–53.

[17] X. Li and C. Qian, "A survey of network function placement," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2016, pp. 948–953.

[18] N. Kazemifard and V. Shah-Mansouri, "Minimum delay function placement and resource allocation for open ran (o-ran) 5g networks," *Computer Networks*, vol. 188, p. 107809, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128621000037

[19] J. M. Ziazet, B. Jaumard, H. Duong, P. Khoshabi, and E. Janulewicz, "A dynamic traffic generator for elastic 5g network slicing," in *2022 IEEE International Symposium on Measurements & Networking (M&N)*, 2022, pp. 1–6.

[20] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167278919305974

[21] L. Li and T. Ye, "Research on throughput prediction of 5g network based on lstm," *Intelligent and Converged Networks*, vol. 3, no. 2, pp. 217–227, 2022.

[22] X. Chang, T. Ji, R. Zhu, Z. Wu, C. Li, and Y. Jiang, "Toward an efficient and dynamic allocation of radio access network slicing resources for 5g era," *IEEE Access*, vol. 11, pp. 95 037–95 050, 2023.