Auto Scaling of Data Plane VNFs in 5G Networks

Tulja Vamshi Kiran Buyakar, Anil Kumar Rangisetti, Antony Franklin A, and Bheemarjuna Reddy Tamma
Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India
Email: [cs16mtech11020, cs12p1001, antony.franklin, and tbr]@iith.ac.in

***Abstract*—In order to meet the traffic demand from diverse next generation wireless network applications and exponentially increasing mobile subscriptions, various 5G network architectures are proposed by leveraging Software Defined Networking (SDN) and Network Function Virtualization (NFV) technologies. Network slicing will be one of the 5G technologies that would support next-generation wireless applications over a shared network infrastructure. However, improper network slicing may lead to either over-provisioning or under-utilization of the underlying network infrastructure resources, especially the 5G core network. Over-provisioning of data plane components such as Serving Gateway (SGW) and Packet Data Network Gateway (PGW) can lead to higher CAPEX and OPEX to mobile operators. In this paper, we propose a novel auto-scaling approach called Bit rate Aware Auto Scaling (BAAS) that maintains a precise UE bit rate requirement in the network slices without over-provisioning of data plane resources.**

I. Introduction

Next generation wireless applications like Internet of Things (IoT), mobile broadband, healthcare applications, *etc.* are demanding a variety of requirements concerning throughput, latency, jitter, and packet losse [1]. To handle these needs, operators are considering Software Defined Networking (SDN), Network Function Virtualization (NFV), and Cloud based platforms [2], [3] for 5G network. These platforms could provide unprecedented freedom in openness, flexibility, and scalability. However, operators cannot utilize the 5G architecture at full extent unless efficient utilization techniques for infrastructure resources [4] are developed. This will result in reduction in Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) for the mobile operators.

Network slicing is a new approach to handle underlying infrastructure efficiently and maximize the utilization of the 5G infrastrucute [5]. Network slicing allows the operators to deploy various application verticals over the 5G network without compromising Quality of Service (QoS) of applications and preventing under-utilization or over-provisioning of resources. It is possible by creating individual end-to-end (E2E) logical network slices for each of the vertical deployed over the 5G network. However, it leads to other issues related to logical isolation of network slices, provisioning for independent management of network slices, ensuring security, and meeting application specific QoS requirements.

As shown in Fig. 1 network slicing based 5G architecture using SDN and NFV technologies is helpful to deploy various next generation wireless applications such as IoT, mobile broadband, and health care [6], [7]. The software-defined and virtualized platform also provide an effective orchestration of

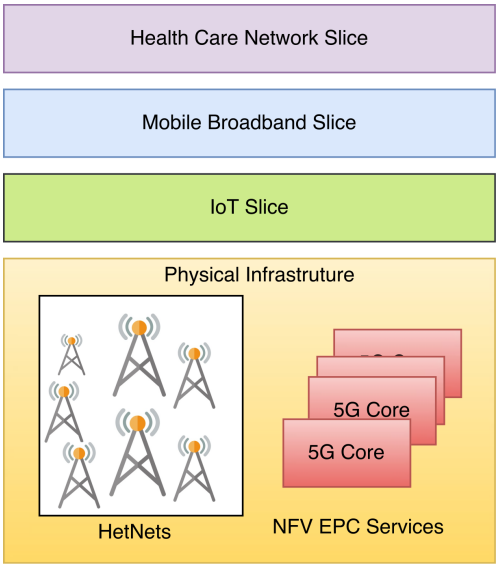

Fig. 1. A Network Slicing 5G Architecture.

E2E network slices. In addition to effectively slicing underlying infrastructure for various applications, there is also need for flexible and scalable resource provisioning mechanism for efficient utilization of the resources. For network slicing based 5G architecture, Evolved Packet Core (EPC) needs to be designed with flexible and scalable NFV platforms with Virtualized Network Functions (VNFs) for Mobility Management Entity (MME), Serving Gateway (SGW), and Packet Data Network Gateway (PGW) running over a cloud platform to handle both control signal overhead and congestion in the data plane (DP). Besides, orchestrator of NFV platform helps in monitoring the resource utilization, fault-tolerance, processing resources isolation, and security provisioning.

Our proposed work mainly deals with creation of various core network slices using NFV based 5G architecture and show how to maintain specifically configured bit rates for each of the network slices using auto-scaling approaches. We also propose an auto-scaling approach for scaling of DP VNFs (S-GW/P-GW) and evaluate its performance in terms of network throughput and number of active DPs needed for handling different traffic loads. However, scaling of resources of RAN and control plane of the 5G core network is not the scope of this work. Our main contributions in this paper are:

- Implementation of a network slicing based 5G network architecture by extending NFV-LTE-EPC [8] framework.
- Proposed a Bit rate Aware Auto Scaling (BAAS) algorithm for auto-scaling of DP NFVs in the 5G core

978-3-901882-98-2 © 2017 IFIP

network.

- Evaluated the network slicing based 5G architecture implementation and BAAS in terms of maintaining the bit rates of DPs and number of DPs required.

II. Related Work

In [7], the authors discussed various coping technologies for efficient utilization of the concept of E2E network slicing in 5G mobile networks. The authors also proposed that a Software Defined Mobile network Control (SDM-C) for E2E flexible network slicing. This work mainly describes a conceptual design of the network slicing in 5G networks. In [6], the authors discussed various fundamental concepts like Network Functions (NFs), infrastructures, virtualization, orchestration, and isolation for designing network slices. Besides, authors also summarized various challenges and benefits of network slicing realization using Open Networking Foundation (ONF) based architectures and NFV based architectures. Finally, SDN and NFV based network slicing architectures are proposed to overcome the challenges and provide maximum flexibility and scaling features to various E2E logical networks realized over 5G network architectures.

In [9], the authors proposed the creation of network slices based on QoS/security service requirements for various service level descriptions. Besides, they proposed a framework and mechanism to enable application services according to its high-level application service provider description. In detail, the framework defines concepts for application-specific slice selection and UEs associations and routing within 5G network. However, authors do not provide any system implementation details and evaluation results.

In [10], the authors developed and evaluated both SDN-based and NFV-based LTE EPC implementations. In their study, various scenarios identified that an SDN-based LTE EPC is well suited for handling large amount of data traffic since it incurs minor overhead for forwarding packets from the kernel or the switching hardware compared to an NFV setup where forwarding decisions are made in user space. On the other hand, an NFV based EPC system is well suitable for handling massive control signal overhead. In this work, we used the NFV-based LTE EPC framework and extended and modified it for our 5G network slicing to evaluate our proposed auto-scaling of DPs.

III. System Model

In our work, we assume that the 5G network slicing architecture can be deployed using an NFV-based LTE system architecture. In the NFV-based LTE-EPC system, major control plane and data plane components such as MME, SGW, and PGW are implemented as VNFs. In our system model, scaling of control plane components and RAN are not discussed as it is outside the scope of this paper. In network slicing system model, scaling of DP is defined as dynamically invoking SGW and PGW VNFs using lightweight virtualization platforms called Linux Containers (LXCs) with required network configuration of interfaces (S1-MME, S1-U, S5/S8 in core network).

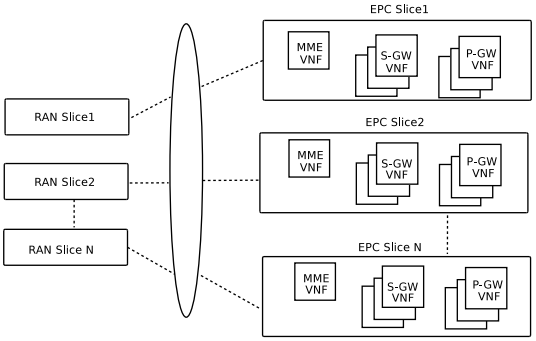

Fig. 2. Network Slicing System Model.

As shown in Fig. 2, network slicing system (separate E2E network slices) is created over the NFV-based LTE-EPC architecture. Specifically, DP VNFs of the network slices are defined to meet specific Maximum DP Throughput (MDT) using processing resources. It means, the DP of a slice can provide only upto a throughput of MDT due to the limitation of the processing resources. To create a DP slice with a given MDT requirement, TCP Socket Buffer Size (TCP-SBSZ) implementation is used [11]. For a DP, TCP-SBSZ of SGW and PGW are set according to its required MDT. In specific, the TCP-SBSZ of SGW and PGW should be sufficiently large so that the flows through the slices can saturate the underlying DP path. When there is no competing traffic, the connections will be able to saturate the path if its TCP sender window is equal to the Bandwidth Delay Product (BDP) of the path. The BDP and its relation to TCP throughput and socket buffer sizing are well studied in the literature [11].

In our model, we use the MME to monitor the traffic loads of the slices and to allocate DPs with SGW and PGW. In each slice, the MME monitors traffic load of DP and initiate auto-scaling of DPs. The MME module is extended with the proposed scaling approach to support auto-scaling of DPs.

IV. Bit rate Aware Auto Scaling (BAAS)

In this paper, we propose a Bit rate Aware Auto Scaling (BAAS) mechanism for auto-scaling of DPs. Timely monitoring of throughput of a DP can be done by employing a network monitoring tool at the SGW VNF. For a network slice, the observed throughput of its DP is reported periodically by the SGW VNF to the MME which will then decide scaling-up (or scaling-down) of DPs for that particular network slice. Initially, a slice starts with single DP to avoid over-provisioning. Then the MME scales-up the DPs of the slice when its throughput utilization (TU) reaches a threshold called TU_{THR} (*e.g.,* 90%). In our auto scaling algorithm TU is the scaling factor to scale-up DPs. Hence, UEs of the slice will not suffer from reduced bit rates. Setting higher TU_{THR} could lead to maximum utilization of the DPs. However, it could result in less available time to setup the DP. Choosing a lower

value for TU_{THR} will lead to wastage of the DP resources. Hence, this value needs to be chosen carefully. We have opted for this as 90% in our experimental evaluation by keeping 10% as the buffer to contain fluctuations in the system load. So, BAAS approach tries to utilize each of DPs up to 90% of TU. For example, over the time, TU of some of the DPs might get lower than TU_{THR}, then BAAS instead of scaling new DPs, it selects a DP with maximum TU ($< TU_{THR}$) and uses it for serving any new UE arrivals.

Besides ensuring MDT requirement of a slice, another objective of BAAS is to minimize resource utilization of the slices without severely affecting throughput of UEs. In order to reduce the overall resource consumption of the slices, auto scale-down of DPs is necessary whenever possible. As scaling down of active DP could result in an overhead of seamlessly transferring the flows of active UEs to other DPs, BAAS opportunistically wait for completion of all UEs of a DP before shutting it down. Therefore, it does not cause any overhead in transferring flows of active UEs.

Algorithm 1 Bit rate Aware Auto Scaling (BAAS)

```
Runs at MME of the Network
Start each slice with single DP to avoid over provisioning
Monitor scaling factor TU of each DP in the slice
for each new UE request do
    TargetDP ← select a DP with maximum TU (TU<TU_THR) in the slice
    Accommodate the UE in TargetDP
    if TU of TargetDP >= TU_THR then
        Scale-up the DP by creating a new DP
        Deploy LXCs of SGW and PGW for the new DP
    end if
end for
Periodically runs auto scale-down:
ScaleDownDP ← select a DP with no active UEs in the slice
Shutdown the ScaleDownDP
```

V. Simulation Setup and Results

To implement the auto scaling of DPs (SGW/PGW) in network slicing based 5G architecture, we have extended the NFV based LTE-EPC implementation from [8]. In general, auto scaling of either SGW or PGW is possible. But in our implementation, scaling of a DP results in scaling of both SGW and PGW, as the PGW in the NFV based LTE-EPC implementation is not handling any particular UE processing function. The NFV based LTE-EPC is extended to create network slices using VNF of MME, SGW, and PGW. The network slicing model used in our implementation is given in Fig. 2 which is realized using lightweight LXCs. Separate LXCs are used for deploying MME, SGW, and PGW. Each DP consist of an SGW and a PGW and each DP of the slice is setup with a specific MDT. MDT of the DP is defined by setting up the TCP-SBSZ of read and write buffers of SGW and PGW. In our implementation, we created four network slices, Slice-1 to Slice-4 with MDT of 25 Mbps, 31 Mbps, 42 Mbps, and 52 Mbps, respectively. Fig. 3 shows various slices in the network and their MDTs.

TABLE I
Simulation Parameters

Parameter	Value
UE arrival	Poisson distribution
UE traffic duration	Uniform distribution (5-20s)
MDT of DP in Slice-2	31.25 Mbps
UE arrival rate during 0-65 secs (λ1)	2
UE arrival rate during 65-130 secs (λ2)	1
UE arrival rate during 130-185 secs (λ3)	6
UE arrival rate during 185-250 secs (λ4)	4
UE traffic	iperf3 TCP
TCP-SBSZ of Slice-2 DP	128 KB
Scaling factor	TU
TU_{THR}	90%
Simulation time	260 secs

The data traffic is generated using the tool "iperf3" that generates TCP traffic with a particular data rate and time duration. The traffic load can be varied by tuning the parameters of "iperf3". In the simulations, DP throughput is measured from the "iperf3" output/statistics. We show that with different traffic load distribution, our proposed BAAS algorithm performs auto scaling of DPs effectively in Slice-2. In our experimental scenario there is enough bandwidth available for all DPs in the system. So the paper did not include results capturing the influence of scaling-up (or scaling-down) of DPs in a slice on other slices in the network. Simulation parameters for traffic load distribution and Slice-2 specific parameters are given in Table I.

Fig. 4 shows the mean arrival rate(λ) of UEs in various time intervals. Fig. 5 describes the number of active UEs present in Slice-2 over the simulation time. Fig. 6 shows the observed network throughput over the simulation time. It is observed that during t=0 sec and t=25 secs, there is a sudden rise in the network throughput due to more UE arrival and it is causing TU of DP crosses TU_{THR} of 90%. Hence, it leads to scale-up of DP by adding a new DP (starting of a VNF of SGW/PGW) to maintain MDT of DPs. Now the number of DPs becomes 2, as observed from the Fig. 7.

After t=100 secs, Slice-2 is introduced with low traffic load by UE arrival rate with λ2=1 and observed that the necessary scale-up and scale-down of DPs is done by BAAS (refer Fig. 7). As shown in Fig. 5, after t=130 secs, the number of active UEs is increased and hence to meet this traffic load of UEs in Slice-2 and maintain MDT of each DP, BAAS started scaling up of DPs to preserve individual MDT of DPs. In Fig. 7, between t=130 secs and t=160 secs, it can be observed that there are new active DPs scaled up and the number of DPs becomes 5. From t=180 secs, as the number of active UEs gradually decreased, DPs with no active UEs are getting shutdown by BAAS. As a result, we are effectively saving processing resource consumption due to unused DPs. In the Fig. 7, after t=200 secs, it can be observed that there are gradual scaling down of DPs. Finally, we summarize the number of active DPs in the network with

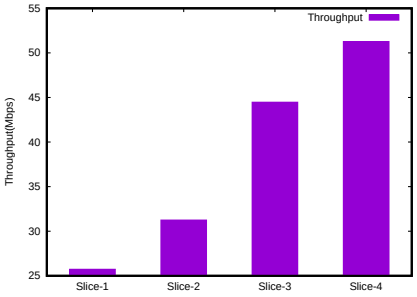

Fig. 3. Maximum throughput observed in various slices.

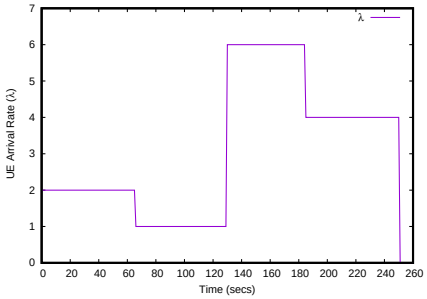

Fig. 4. UE arrival rates (λ) at various time intervals in the network.

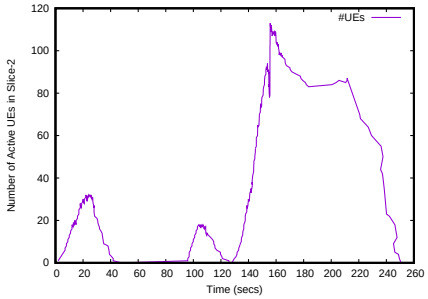

Fig. 5. Number of active UEs in Slice-2 with time.

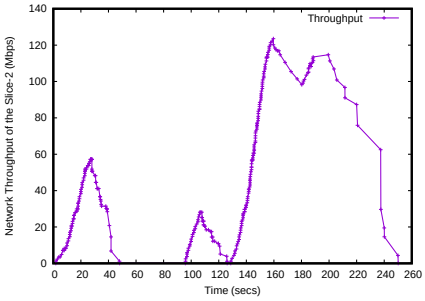

Fig. 6. Network throughput of Slice-2 with time.

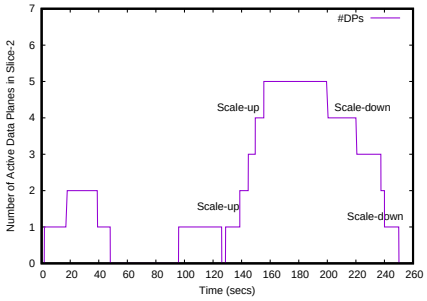

Fig. 7. Number of active DPs in Slice-2 with time.

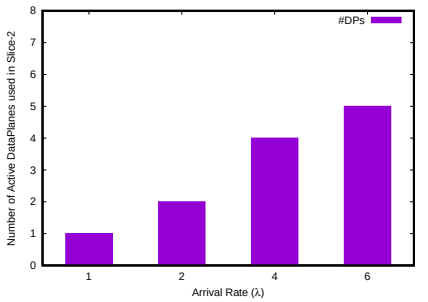

Fig. 8. Number of active DPs for different arrival rates.

the given arrival pattern of UEs in Fig. 8. It is clear that BAAS dynamically vary the number of active DPs based on the load in the network. Without BAAS mechanism the network would either suffer from reduced bit rates or over-provisioning of underlying resource.

VI. Conclusions and Future Work

In this work, we demonstrated the network slicing based 5G architecture using LXCs. To the best of our knowledge, this is the first attempt to show a real feasibility of network slicing based 5G system. As it is first attempt to implement a network slicing based 5G system, we have provided only the proof of concept result in this paper. We also proposed an auto-scaling approach called Bit rate Aware Auto Scaling (BAAS) for maintaining bit rates of DPs and avoiding over-provisioning of underlying DP VNFs for network slices. Unlike general auto-scaling approaches which are based on processing resources utilization, BAAS builds on the actual throughput utilization of DPs. Hence, it tries to scale-up new VNFs of DPs to maintain MDT. This is our first attempt to create a network slicing architecture for 5G. We plan to extend our work to create network slicing architecture with various verticals and test the system with the different auto-scaling mechanisms.

ACKNOWLEDGMENT

This work was supported by the research project "Low Latency Network Architecture and Protocols for 5G Systems and IoT" funded by Science and Engineering Research Board (SERB), Government of India.

References

[1] "Cisco Visual Networking Index, Global mobile data traffic forecast update, 2015–2020," *white paper*, 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[2] A. Brunstrom, K.-J. Grinnemo, J. Taheri *et al.*, "SDN/NFV-based mobile packet core network architectures: A survey," *IEEE Communications Surveys & Tutorials*, 2017.

[3] C. Bouras, A. Kollia, and A. Papazois, "SDN & NFV in 5G: Advancements and challenges," in *Proc. of Innovations in Clouds, Internet and Networks (ICIN)*. IEEE, 2017, pp. 107–111.

[4] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[5] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[6] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[7] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[8] "NFV-LTE-EPC," https://github.com/networkedsystemsIITB.

[9] V. K. Choyi, A. Abdel-Hamid, Y. Shah, S. Ferdi, and A. Brusilovsky, "Network slice selection, assignment and routing within 5G networks," in *Proc. of IEEE Conference on Standards for Communications and Networking (CSCN),*, 2016, pp. 1–7.

[10] A. Jain, N. Sadagopan, S. K. Lohani, and M. Vutukuru, "A comparison of SDN and NFV for re-designing the LTE packet core," in *Proc. of Network Function Virtualization and Software Defined Networks (NFV-SDN)*. IEEE, 2016, pp. 74–80.

[11] R. S. Prasad, M. Jain, and C. Dovrolis, "Socket buffer auto-sizing for high-performance data transfers," *Journal of GRID computing*, vol. 1, no. 4, pp. 361–376, 2003.