Reliability-Aware Network Slicing in Elastic Demand Scenarios

Rafael L. Gomes, *Member, IEEE* Luiz F. Bittencourt, *Member, IEEE* and Edmundo R. M. Madeira, *Member, IEEE*

Abstract—**The Internet is an essential tool for the society as a whole, being the basis for several services. This importance increased the requirements for Internet Service Providers (ISPs). The current Internet infrastructure is limited, which often compromises the Quality of Service (QoS) and Quality of Experience (QoE) of users. Therefore, ISPs need to evolve their technologies and management capacity. One key approach is network slicing, which allows the management of reliability and elastic resource demand. This article discusses the reliability requirements of the network slicing process as well as the features and challenges of elastic demand scenarios. Additionally, it presents a reliability strategy for network slicing. The results from the experiments performed, using a dataset with real network demands, suggest that the reliability strategy mitigates the impact of physical failures over the Internet access service.**

I. INTRODUCTION

The rise of infrastructures based on Software Defined Networks and Virtualization enabled effective and operative network and service deployments on top of network slices. As a result, a new scenario where Internet Service Providers may offer not only Internet Access Services (IASs), but also virtualized computational capabilities by elastically dividing the network infrastructure into network slices offering adaptable network functions and services [1].

Internet Service Providers (ISPs) offer services through a Service Level Agreement (SLA), where client requirements and guarantees are defined. Service delivery can be impacted by the elastic demand for network resources, caused by several factors of the modern society, such as human mobility within cities and heterogeneity of behavior inside the same local network [2]. In an elastic demand scenario, the generated traffic volume varies according to the user's context and time periods [3]. When not treated properly, this elastic demand may result in Quality of Service (QoS) and/or Quality of Experience (QoE) degradation (slowness and connection interruption) and wastage of resources. Consequently, clients are not satisfied with the service, and providers can face increased Capital expenditure (CAPEX) and Operational expenditure (OPEX).

Reliability is a key requirement in SLAs, since it defines service availability: reliability represents the probability of a service to remain operational even when failures occur in the physical network [4]. Reliability in elastic demand scenarios brings dynamic requirements for ISPs, increasing complexity of network slices management. Addressing these requirements involves not only reactive actions, but also strategic planning of the network infrastructure configuration. Designing solutions to mitigate potential impacts on clients QoS/QoE in face of network failures and elastic demand is necessary.

In this context, this article discusses the reliability requirements of the network slicing process and the challenges in elastic demand scenarios. The characteristics of network slices and the network slice life-cycle management are presented, and their impact over service delivery is discussed. Additionally, this article presents a reliability strategy to evolve the approach previously proposed by the authors [5]. The original algorithm coupled reliability constraints with the search for disjoint paths in the allocation process.

The proposed reliability strategy for network slicing is applied in conjunction with slicing allocation algorithms to define alternative connectivity paths based on a link importance selection, encouraging the usage of low saturated links that are part of a possible solution. The aim is to combine efficient slice allocation and network reliability in a flexible manner. Thus, the reliability strategy proposed in this article can be attached to different slicing allocation algorithms.

R. L. Gomes. is with the State University of Ceara (UECE).

L. F. Bittencourt and E. R. M. Madeira are with the University of Campinas (UNICAMP).

The performance of the proposed reliability strategy was evaluated using a dataset with real network demands [6], instead of synthetic data used in [5]. The results suggest that the reliability strategy mitigates the impact of physical network failures on the internet access service.

The next section details the reliability requirements in the network slicing process. Section III describes the existing related work. Section IV presents the designed reliability strategy, while Section V describes a case study and discusses results. Section VI presents some challenges and Section VII concludes the paper.

II. RELIABILITY REQUIREMENTS OF SLICING

New services based on Software Defined Networks (SDN), Network Function Virtualization (NFV), and Network Slicing are becoming essential for the ISPs, where continuity is one of the most essential features, i.e., high availability services that are resilient to possible failures and other disruptions. Service continuity is directly related to customer expectations and reliability requirements reflected in the SLA (the *desired reliability*). A reliable network service should provide availability by treating failures, but also provide consistent quality of service by ensuring requirements (e.g., desired bandwidth) are followed. One way to provide reliability is to ensure adaptive resource provisioning and strategic planning for dynamic situations resulting from failures or variable requirements.

Failure of network elements occurs randomly at any time and place, generating unusual situations. Introducing proper planning and allocation of resources that take failures and changing networking conditions into consideration enables proper utilization of networking resources and can provide service continuity and quality requirements fulfillment, thus creating a reliable service offering.

Considering the discussion above, reliability must be taken into account when designing and running network services. Typically, efforts towards reliability assurance come after services deployment. Notwithstanding, reliability and service deployment should be considered altogether, since failures and disruptions can impact overall service operation.

III. RELATED WORK

We discuss existing literature focusing on network slices, reliability, and approaches to deal with the elastic demand problem in service delivery.

Chiha et al. [7] design a model that computes slices cost to enable proper pricing of end-to-end services based on SDN and NFV. The authors focus on the network infrastructure provider perspective since techno-economic analysis are considered. The proposed model acts in conjunction with a network function (NF) dimensioning model to achieve the desired evaluation of costs.

D. Gutierrez-Estevez et al. [3] propose an architecture with built-in Artificial Intelligence features that allow the exploitation of elasticity. The authors describe a taxonomy of learning mechanisms for network elasticity. This architecture enables the resource elasticity as a key factor to make efficient use of the computational resources.

Caballero et al. [8] design a slicing framework using a game-theoretical approach to reach Nash Equilibrium with a user dropping method and a resource allocation scheme. In conjunction with an admission control policy, which is used to satisfy the rate requirement of users, the framework can guarantee the requirements of all users.

P. Mohan et al. [9] describe a reliable slice embedding scheme to reduce the number of service chains affected when physical machines fail. The authors propose an optimization formulation to instantiate virtual network functions (VNFs) such that the number of affected service chains are minimized while routing and VNF placement constraints are satisfied in conjunction with clients requirements.

V. Petrov et al. [4] introduce a softwarized framework for reliability of mission-critical traffic. The authors also present a mathematical model in softwarized networks to define the process of critical session transfers as well as the impact of these transfers over sessions from other users.

The related work discussed above does not present solutions for bringing reliability into slices allocation considering elastic scenarios. Combining efficient allocation in an elastic scenario where demand varies through time with reliability requirements is still an open issue.

IV. RELIABILITY STRATEGY FOR NETWORK SLICING

Motivated by the frustration that Internet users experience when facing problems such as frequent disconnections, slowness, and large delays, we present a reliability strategy to be applied during slice allocation considering an elastic scenario with varying demands throughout the day. The goal is to avoid low reliability slices in service delivery while also being aware of different demands during a certain period of time. As a result, reliable quality of service can be offered by ISPs to their clients, while ensuring a proper utilization of network resources.

Usually, reliability is expressed as a value in the $[0, 1]$ interval, with larger values representing higher reliability. This value describes the probability/capacity of the network links that support the service delivery to be operational [4]. The desired reliability (i.e., the minimum value to be achieved) is defined in the SLA. Thus, reliability is not only concerned with reactive actions that address and/or try to contour post-failure impacts, but it should also be concerned with pre-failure strategic planning. In our context, a failure can be associated with network equipment malfunction or other events that run into violation of QoS constraints. For example, a failure can be a switch port that stops working and shutdowns a link, or a switch port that gets overloaded by traffic in a way that impedes slice traffic to be forwarded obeying QoS constraints. In the occurrence of such events, the network should be prepared to keep clients' QoS.

As demands are variable, ISPs should define slices according to different periods of the day. Network slices can support service delivery over the Internet by providing a dynamically and independently managed network, resulting in a logically independent network, on top of physical or virtual networks. A network slice is defined as a set of components that will compose the slice, which are chosen by a resource allocation algorithm that optimizes network usage following requirements of clients throughout the day.

Figure 1 illustrates the scenario depicted above. In the morning (Time 1), the slice has network resources allocated to it such that the client requirement is met. Later in the day, the bandwidth demands change and the slice is adjusted (Time 2). Similarly, this situation occurs between Times 2 and 3. Hence, the set of network elements allocated to the example slice at Time 1 may be different from Time 2 and Time 3.

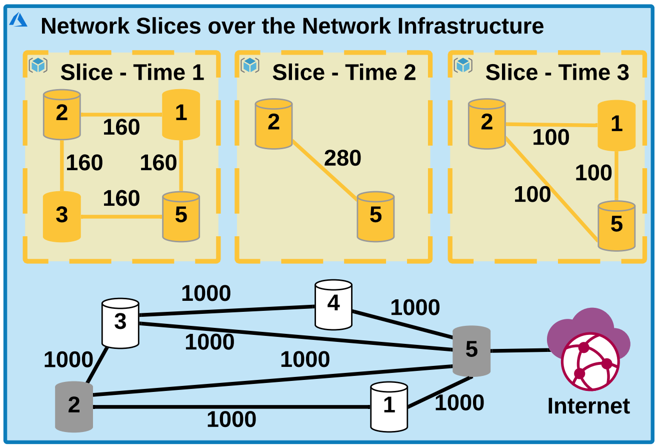

Fig. 1. Network Slices Allocation Process with three periods of time requesting 320Mbps, 280Mbps and 200Mbps, respectively.

According to [1], two of the most important steps in deploying a network slice are *Preparation* and *Configuration*. The *Preparation* plans the network slice and defines the set of physical network elements (links and switches) that will be part of a slice, which are then allocated during the *Configuration*. The *Configuration* step applies the network slice designed over the network infrastructure, interacting with the network hypervisor of the network infrastructure. Thus, these two steps control the lifecycle of a network slice, composing the strategic planning of the network environment. In this way, the allocation algorithm with the reliability strategy is executed inside the *Preparation* step.

The algorithm used to allocate network resources for slices directly impacts the reliability and the bandwidth availability of the infrastructure. Using longer connectivity paths (with more links) consumes more aggregated bandwidth of the infrastructure and reduces reliability, as the more devices in a path, the larger the aggregated failure probability. However, always allocating the same link to reduce path length for most slices will saturate its bandwidth quickly. This can compromise future requests that need that link to be operational.

In this context, the proposed reliability strategy is designed to be applied in conjunction with an allocation algorithm, as illustrated in Figure 2. The reliability strategy performs a relative disjoint path search, looking for alternative paths until the desired reliability for the slice is reached. Simultaneously, the proposed strategy defines these alternative connectivity paths based on a link importance selection. Thus, the proposed reliability strategy is based on two central definitions:

- Link Suitability: Several criteria can be used to evaluate the suitability of a link to be part

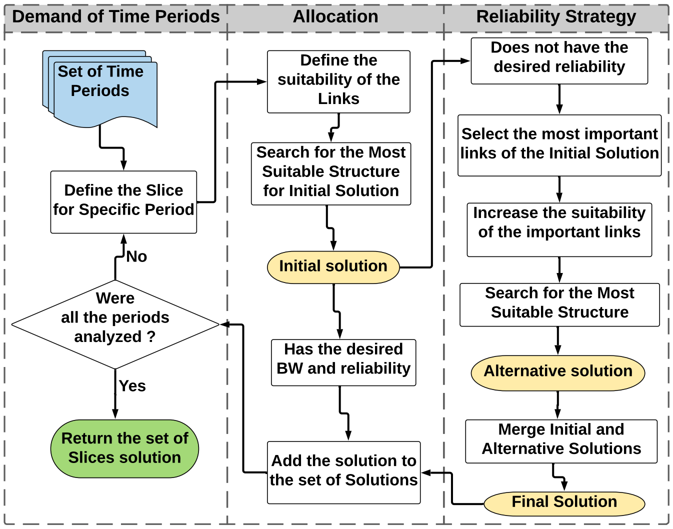

Fig. 2. Overview of Reliability strategy.

of a network slice. One possible approach is to measure how the infrastructure is impacted by a slice allocation. For example, a link may be considered more suitable if it has the highest amount of bandwidth available, as this approach can help in avoiding premature saturation of links while it allows to increase the allocated bandwidth if the next time period requires more bandwidth for the slice. As a consequence, it can also maximize the use of network resources by allowing more active customers (i.e., users with SLAs requirements fulfilled).

- Link Importance: Regarding the identification of the most important links, the reliability strategy considers a link as important in the following situations: (i) during the initial search process, it is part of one of the several possible interconnection paths between source and destination; and, (ii) it will not be saturated if the desired bandwidth is allocated. When a link is considered important, its suitability is increased (in relation to the suitability defined initially in the allocation algorithm), which encourages its allocation. At the same time, the allocation of links considered not important is discouraged, forcing a new search process to find alternative paths using the important links as a basis. Unlike the previous work [5], which does not apply any importance criterion, the reliability strategy in this article evolved the survivability capacity of the network slices by identifying the important links of the initial solution.

Recall slice structures represent the network physical elements (switches and links) that will be part of the network slice and, consequently, have their resources allocated. The input for the slice allocation process is a set of network resources demand (such as bandwidth), split into time periods, and the desired reliability to be achieved. These time periods are defined in the SLA between the client and the service provider, where the slice allocation process needs to define a set of slice structures to be deployed at each time period defined. The most appropriate set of time periods can be defined using several approaches, such as forecasting and optimization (as a third-party service), explicit requests by the client, as an ISP offering (according to the available resources), among others. Thus, the reliability strategy is independent of the time periods definition approach utilized.

The reliability strategy execution, in conjunction with a generic slice allocation algorithm, is represented in the flowchart of Figure 2. Initially, the ISP receives the set of time periods and their requirements, then it starts the process to define a slice for a specific period of time as follows. Initially, the suitability of the links for the current network slice processing is computed; then, the search for the *initial solution* with the most suitable interconnection path is performed. If the *initial solution* does not meet the desired bandwidth, the slice definition process fails, since no possible solution was found. When the *initial solution* has the desired bandwidth, its reliability is evaluated. If it has the required reliability, then the *initial solution* is included as the solution for this period of time; otherwise, the reliability strategy is invoked to find an *alternative solution*.

First, the reliability strategy identifies the most important links of the *initial solution* and encourages their usage by increasing their suitability. Then, it searches for an *alternative solution* (i.e., an alternative interconnection path) considering the updated links suitability. Next, the reliability strategy merges the *initial* and *alternative* solutions, creating the *final solution*. Lastly, if the *final solution* has the desired reliability, then it is included as the solution for that period of time; otherwise, the slice definition process fails. An example of the execution of the slice definition process for a period of time is illustrated in Figure 3 (which aims to interconnect nodes 3 and 7): First, the suitability of the links is calculated; then, the *Initial* solution is processed; if it does not reach the desired reliability, the link importance is calculated, and the *alternative* solution is processed (considering the node 2 as part of the solution instead of node 6); Lastly, the final structure is defined, merging both *initial* and *alternative* solutions. During the example, the terms "high", "medium" or "low" are an illustrative representation for the suitability according to the links' available bandwidth in relation to the bandwidth requested in the SLA, as described previously. This representation is used to ease the understanding of the proposed strategy in the example described.

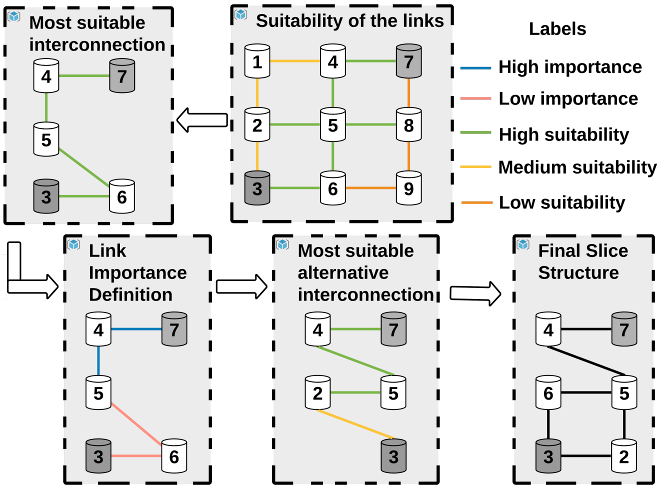

Fig. 3. Example of slice structure solution for a period of time.

The idea of the reliability strategy in the *alternative* solution is to avoid the non-essential links of the *initial* solution while keeping the most important links (encouraging allocation by increasing their suitability). The desired bandwidth is allocated in full in the links present in both solutions (*initial* and *alternative*), while the other nodes that have redundancy allocate half of the network resources, since the aggregated allocation meets the requirements.

It is also necessary to consider the possible overhead on the performance of the network infrastructure in front of the changes in the structure of the slices from one time period to another [10]. In the proposed reliability strategy, this impact can be mitigated by the definition of link suitability, since it allows the allocation process to smooth the necessary modifications and maintain the slice structure as long as it reaches the desired reliability (defined in the SLA).

Finally, the proposed strategy takes the computed reliability as a criterion to constraint the slice suitability. Note that different reliability calculation methods can be applied, since the proposed strategy is agnostic of the reliability calculation method. Currently, we focused on link failures for the reliability calculation. Nevertheless, using the link and node failure approach demands new factors to be considered during the link importance definition, since it brings correlated failures that impact the alternative paths deployed.

V. CASE STUDY

This section shows a case study to evaluate the proposed reliability strategy related to the network slices allocation process. The experiments used a real network traffic dataset collected in the backbone of the State University of Ceará (UECE) [6], [11].

One hundred sets of slice requests were randomly generated, where each set was composed of one hundred requests considering the nodes to be interconnected (randomly chosen from a uniform distribution). These nodes represent the client's border gateway to the Internet. Three metrics are evaluated: (A) Number of network slices that provide the requested bandwidth and connectivity (i.e., operational slices); (B) Number of saturated links; and (C) Running time.

Network slices were allocated in the GEANT (39 nodes and 59 links) and ATT (24 nodes and 57 links) network topologies [12]. These two network topologies were selected due to their centrality level and robust structure, enhancing the necessity of suitable approaches for slices deployment over them [12].

We compare the performance of existing algorithms, namely REENC [5] and PETIC [2], with and without the proposed reliability strategy. In summary, PETIC individually defines the network slices for each period of the day, while REENC tries to reuse the structure of the slice from the previous time period to define the slice for the current time period. Details about the algorithms and the experimental setup can be found in [5], [2].

Figure 4 illustrates the number of operational slices after the occurrence of failures in the network infrastructure. In general, active slices depict the ability of the reliability strategy to maintain service delivery. The "x" axis in Figure 4 shows the percentage of network components that randomly fail: “0%” illustrates the situation where the network infrastructure is fully operational, “5%” when 5% of the network infrastructure fails, and so on. On the other hand, the “y” axis represents the number of allocations fulfilling the SLA, which considers reliability and bandwidth.

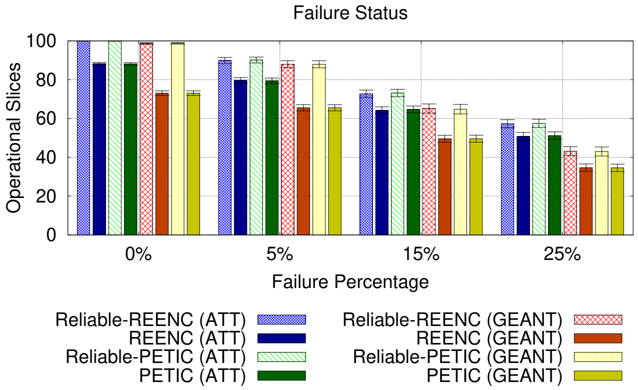

Fig. 4. Post-failure status.

Figure 4 shows the proposed strategy achieves a good performance for reliable slices definition and keeps the allocated slices operational, maximizing the survivability of the service delivery regardless the algorithm applied. Also, not using the reliability strategy results in a worse performance of the slice allocation in GEANT compared to ATT due to the lower centrality of the GEANT topology [12]. The reliability strategy use equates the performance of the slice allocation in both topologies, since it explores their redundancy capacity.

As described in the previous section, the incremental path redundancy approach of the proposed reliability strategy tends to include more network elements in the network slice structure. Thus, the analysis of the impact over the bandwidth availability is crucial. In this article we considered a link as saturated whenever it has available less than 30% of its original bandwidth, since this low availability will directly affect the capacity of the network to deploy recovery policies. The proposed reliability strategy results in a small number of saturated links after all requests were analyzed: considering the ATT topology, it has 57 links and the reliability strategy saturated less than 3% of the links in the worst case. Naturally, the higher the number of successfully allocated slices, the higher the network resource usage.

An existing drawback of introducing a reliability strategy is the slice allocation algorithm running time, since it adds new steps during the slicing allocation, increasing its time complexity. In the experiments performed, the execution time was, in average, 30% higher: 39 seconds with the reliability strategy and 30 seconds without it. On the other hand, it resulted in 15% more requests solved. As described in Section IV, the reliability strategy algorithm is invoked only in cases where the desired reliability was not achieved in the initial search. In this way, the extra processing steps are executed only to improve the slice allocation process to fulfill the requirements defined in the SLA.

The results suggest the reliability strategy is capable of improving the slice allocation with (i) low impact in the available bandwidth of the network infrastructure; and (ii) definition of reliable network slices according to the periods of time, mitigating the impact of failures in the network. Regarding the allocation algorithms, when the reliability strategy is applied, the REENC presents a better performance when the percentage of failures increases. This occurs due to REENC’s tendency to keep the slice structure when it reaches the reliability of the previous time period, reducing the failure points possibilities from one time period to another one.

VI. CHALLENGES

Challenges in network slicing with dynamic demand appear as the optimization of resource allocation, which is a time-consuming task as also is the reconfiguration of the network to deploy those slices. Overcoming those challenges involves defining meaningful periods of time that capture overall client requirements dynamics in which slices optimization and reconfiguration can be performed on time. While too short periods of time may impose excessive reconfigurations, which overload the system and reduce availability, too long periods of time may result in under/over utilization of super/sub provisioned slices. Therefore, defining the ideal time slot sizes for different scenarios to achieve a balance between reconfigurations and utilization is a core issue for the performance of the sliced network.

An ideal time slot size will depend on the traffic behavior and/or SLAs defined. Traffic characterization can rely on monitoring tools and pattern analysis, e.g., using machine learning techniques or time series modeling. Approaches, such as [13], can be used to model the client’s traffic and define the optimal set of time periods. For example, the output of such analysis can be used to feed the network management system with predicted traffic behavior in order to anticipate adaptation needs and precompute the network resource allocation into slices in the next time period.

VII. CONCLUSION

Elastic demand for network resources impacts quality of service, where it is necessary to dynamically adapt the resources allocated. A key process to deal with elastic services is the preparation step in the life-cycle of a network slice, which defines the network elements (links and nodes) that will compose the slice. Strategies must be designed to achieve the services requirements, among which reliability is one of the most important. This article discussed reliability in network slicing deployment and the characteristics of elastic demand that influence the network slices life-cycle. A reliability strategy to network slice preparation for elastic demand scenarios is proposed, and experiments using real resources demand dataset suggest that it improves resilience in service delivery.

ACKNOWLEDGMENT

This work was partially supported by CNPq and INCT of the Future Internet for Smart Cities (CNPq 465446/2014-0, CAPES - Finance Code 001, and FAPESP 2014/50937-1 and 2015/24485-9).

REFERENCES

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[2] R. L. Gomes, F. da Ponte, A. Urbano, L. F. Bittencourt, and E. Madeira, "Strategies for daytime slicing in future internet service providers," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 1, p. e3727, 2020.

[3] D. M. Gutierrez-Estevez, M. Gramaglia, A. D. Domenico, G. Dandachi, S. Khatibi, D. Tsolkas, I. Balan, A. Garcia-Saavedra, U. Elzur, and Y. Wang, "Artificial intelligence for elastic management and orchestration of 5g networks," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 134–141, 2019.

[4] V. Petrov, M. A. Lema, M. Gapeyenko, K. Antonakoglou, D. Moltchanov, F. Sardis, A. Samuylov, S. Andreev, Y. Koucheryavy, and M. Dohler, "Achieving end-to-end reliability of mission-critical traffic in softwarized 5g networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 485–501, 2018.

[5] G. da Silva, D. Oliveira, R. L. Gomes, L. F. Bittencourt, and E. R. M. Madeira, "Reliable network slices based on elastic network resource demand," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, 2020, pp. 1–9.

[6] R. L. Gomes, "Bandwidth usage of university campus," 2020, doi: 10.21227/jw40-y336. IEEE Dataport. [Online]. Available: http://dx.doi.org/10.21227/jw40-y336

[7] A. Chiha, M. Van der Wee, D. Colle, and S. Verbrugge, "Network slicing cost allocation model," *Journal of Network and Systems Management*, pp. 1–33, 2020.

[8] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6419–6432, Oct 2018.

[9] P. M. Mohan and M. Gurusamy, "Resilient vnf placement for service chain embedding in diversified 5g network slices," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[10] 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, May 2017.

[11] R. L. Gomes, "Slicing elastic demand," Accessed on: September 2020. [Online]. Available: https://github.com/rafaellgom/slicing-elastic-demand

[12] D. F. Rueda, E. Calle, and J. L. Marzo, "Robustness comparison of 15 real telecommunication networks: Structural and centrality measurements," *Journal of Network and Systems Management*, vol. 25, no. 2, pp. 269–289, 2017.

[13] E. Harstead and R. Sharpe, "Forecasting of access network bandwidth demands for aggregated subscribers using monte carlo methods," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 199–207, 2015.

Rafael L. Gomes is an Associate Professor at the State University of Ceara (UECE). He received a Ph.D degree in Computer Science from the University of Campinas (UNICAMP) in Brazil. He was a research visitor at Network Research Lab from UCLA in 2014. He is part of technical program committees of several conferences, such as IEEE/IFIP IM, IEEE/IFIP NOMS, IEEE LATINCOM, and others. He has researches on the following topics: SDN, Resilience, and IoT.

Luiz F. Bittencourt is an Associate Professor at the UNICAMP, Brazil. Luiz was awarded with the IEEE Communications Society Latin America Young Professional Award in 2013. He was a visiting researcher in the University of Manchester, Cardiff, UK, and Rutgers University, USA. He serves as associate editor for the IEEE Cloud Computing Magazine, the Computers and Electrical Engineering journal, and the Internet of Things Journal. His interests are in the areas of network management and task schedulers in cloud-fog.

Edmundo R. M. Madeira received a Ph.D. degree in Electrical Engineering from the UNICAMP, Brazil, in 1991. He is a Full Professor with the Institute of Computing at the University of Campinas. He is a member of the editorial board of the Journal of Network and Systems Management (JNSM), Springer. His research interests include network management, future Internet, 5G networks, vehicular networks, and cloud and edge computing.