A Utility-Driven Multi-Queue Admission Control Solution for Network Slicing

Bin Han, Vincenzo Sciancalepore, Di Feng, Xavier Costa-Perez and Hans D. Schotten

A review by Sarthak Garg

1 Introduction

Network Slicing is an emerging 5G technology that allows infrastructure providers to offer "slices" of resources (computational, storage and networking) to network tenants. Advanced admission control policies are required as tenants compete for a limited bunch of available resources. This paper proposes a new slice queuing solution based on a stochastic model for network slicing using a multi-queuing system. The model supports various kinds of slices requests, while enforcing performance metrics and ensuring maximisation of resources.

2 Contribution

The paper starts by mathematically modelling the resource pool and the slice types. Several vectors are defined which represent the resource pool, set of network slices, the resources required to maintain a slice of a certain type, and the assigned sources. Constraints are introduced to ensure *resource feasibility*. It is assumed that certain number of tenants are randomly generating network slice requests which are independent and identically distributed (i.i.d.). Because of a temporary shortage of available resources the tenant might not be able to obtain the requested slice immediately. And so the tenant needs to request the slice again after a certain delay. Both random delay and queuing are discussed and it is concluded that using **multi-level queues** to store delayed slice requests is better since it provides better efficiency and control over resources.
The next section in the paper discusses the types of queue to use for the model. A simple case study is presented which gives a very clear example of how *heterogeneous multi-queuing schemes* suit the purpose better than single-queue and the homogeneous multi-queue schemes. Heterogeneous in the sense that each queue holds a different slice type. The authors decide to use a certain number of *First Come First Serve(FCFS)* queues. A preference matrix is characterised which will be used to decide between the various types of slices, hereby setting the admission strategy. The proper algorithm is provided and an elaborate proof is provided as to why the acceptance of requests in queues is a Poisson process. However, this proof and algorithm has some **serious assumptions** about the

inter arrival times and Independence of requests. Various mathematical concepts related to birth-death Markov processes are discussed, which will be used in the future for performance metrics. To handle impatient tenants, two different models are introduced i.e *Balking model and Reneging model.* Then using detailed mathematical calculations involving queuing theory and probability, the expected waiting time of requests in queues are calculated.
By converging utility rate series and limiting conditional probability functions of the states of the Markov chain, the acceptance rate is maximised and the strategy is optimised. The final part of the paper uses numerical simulations to verify the model characteristics and resource utilisation. 10,000 random cases are generated and the proposed model is compared with existing solutions on these cases. Various charts and plots show that the proposed solution outperforms on various metrics such as *overall network utility rate, the admission rate and the average request waiting time.*

3 Discussion

The context of the paper is **extremely relevant** since it addresses a key problem of today: network resource management and maximising the efficiency and service quality. In the age when 5G network is being rolled out and every industry is starting to become heavily dependent on networks and data usage, this field of research is very beneficial to both the network operators as well as the consumers of various services.
The authors do not propose an entirely revolutionary and original idea but work on an existing heavily researched area while implementing certain new mechanisms- such as using multi level heterogeneous queues instead of single or multi-level homogeneous queues. **The paper heavily relies on mathematics** since it is based on a proposed solution to upcoming 5G technology. The mathematical modelling of the problem is easy to understand and the authors used simple examples and case studies to justify assumptions and decisions which made it quite interesting. It isn't just filled with only theory but also has many practical examples and simulations. The results of the simulations are displayed using intuitive plots and charts which makes them easier to understand.
I felt too many models and tenant request cases were introduced at the same time and the authors should have progressively carried out simulations and metrics on each case one by one. Moreover, there are many instances where **mathematical assumptions are made without any proper justification**. For example, it was assumed that the resource pool is static whereas in reality the resource pool is dynamic with respect to time. Also, it was assumed that the inter arrival time of requests is independent and follows Poisson distribution. There was no suggestion on the corrections to the model in case these assumptions are not valid.
Overall, it can be said that the context of the paper was extremely relevant and the methodology had a fine balance between theory and actual simulations. A significant amount of probability and stochastic process concepts are utilised. The authors provide exhaustive feedback about their positive results. This research paper is perfect to gain insight about queuing theory in computer networks.