JOURNAL OF SCIENCE
SAKARYA UNIVERSITY

Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |

http://www.saujs.sakarya.edu.tr/

Title: Exploring Analytical Model to Performance Optimization for Mobile Application Using End-to-End Network Slicing in Cloud-Based Vehicular Networks

Authors: Yonal Kırsal

Recieved: 2018-03-12 13:19:11

Revised: 2018-05-26 19:09:52

Accepted: 2018-07-31 15:08:44

Article Type: Research Article

Volume: 23

Issue: 1

Month: February

Year: 2019

Pages: 23-34

How to cite
Yonal Kırsal; (2019), Exploring Analytical Model to Performance Optimization for Mobile Application Using End-to-End Network Slicing in Cloud-Based Vehicular Networks. Sakarya University Journal of Science, 23(1), 23-34, DOI: 10.16984/saufenbilder.404414
Access link
http://www.saujs.sakarya.edu.tr/issue/38708/404414

Sakarya University Journal of Science 23(1), 23-34, 2019

Exploring Analytical Model to Performance Optimization for Mobile Application Using End-to-End Network Slicing in Cloud-Based Vehicular Networks

Yönal Kırsal*[1]

ABSTRACT

Today, many large usages of cloud-based vehicular networks and applications have rapidly increased. This rapid increase causes the requirement of systems to be reliable to share their resources without delay in order to ensure a better quality of service (QoS) for mobile users. Hence, network slicing is considered one of the key concepts to enhance QoS in 5G networks. At present, new architectures attempt to provide support for end-to-end server quality mechanisms. A key mechanism of network slicing supported by such modern architectures is able to either handover to better network or migrate services closer to the users as they move around. This can be done by advanced handover and server localization techniques. These sorts of advanced handover and server localization help to maintain the QoS for mobile application in heterogeneous environments. In order to obtain QoS measurements and get the network conditions in a specific area, a cloud-based vehicular network slicing management framework is proposed using an analytical modeling approach. The analytical model results obtained consider real scenarios from the Middlesex University vehicular ad-hoc networks (VANET) testbed. Using this framework, the mobile users will make a decision on which situation is better suited to obtain the service based on the latencies as well as queuing capacities of the networks.

Keywords: network slicing, QoS management, analytical modelling, advanced handover, server localisation, vehicular ad-hoc network

1. INTRODUCTION

The next generation wireless network (e.g., 5G) will need to support new demands from a wide variety of users, machines, industries, governments and other organizations. 5G has several heterogeneous requirements from the application perspective, as well as from an architectural perspective [1,2]. From an application perspective, several Internet of Things (IoT) applications present many strengths and challenges, in terms of power, scalability and latency requirements. At the same time, 5G also needs to be highly secure whilst supporting mobility. Mobile subscribers demand uninterrupted connection from anywhere and anytime as the mobile users move between networks [3]. From an architectural point of view, the new top-down network architecture should be service centric. They fundamentally expose software-defined networking (SDN) and network function virtualization (NFV), where both connectivity services and highly contextualized services should be provided to end users [4]. Hence, in order to meet these requirements, several emerging architectures are under consideration. One of the most important points is network slicing, which allows you to slice the network in a top-down application manner [5]. Efficient network slicing can be achieved by advanced handover and server localization techniques [6,7]. Handovers can be classified according to general characteristics such as network-based/client-based and hard/soft

* Corresponding Author
1 European University of Lefke, Faculty of Engineering, Department of Electronics and Communication Engineering, Lefke, North Cyprus TR-10 Mersin, Turkey, E-mail:ykirsal@eul.edu.tr

handovers. Handovers can be further classified as advanced handover, dealing with the different mechanisms and inputs [8]. In order to improve the QoS, the mobile users may change their network operator based on the level of performance, or are forced to change the network point of attachment (PoA) [8,9,10]. Proactive handover is a type of advanced handover; and attempts to get the condition of the available networks in close proximity at a specific location before the handovers occur. Mobile users can calculate the time before vertical handover (TBVH) [11], using proactive handover policies [12]. Knowledge-based and model-based handovers policies are two types of proactive handover currently being developed. The knowledge-based policies allow mobile users to measure the signal strength of available networks over a given area beforehand [13]. Model-based handover is based on a mathematical model which calculates the point when a vertical handover should occur, as well as the time that mobile users would take to reach that point based on the mobile users' speed and direction. In this paper, the proposed model is a type of model-based handover. A vehicular cloud-based networks-service, delivery framework for video streaming is presented, which considers the network-slicing factor. We consider, services run on localised clouds, specifically Middlesex University cloud, depending on service demands and networks. A mathematical model is developed and a common scenario is also considered between the road-side units (RSUs) for enabling a 5G system with end-to-end (E2E) network slicing.

The rest of the paper is organised as follows: Section II presents the related studies for network slicing on cloud-based vehicular environments. The proposed cloud-based vehicular system is given in Section III. Section IV introduces the analytical model and a solution approach. Section V presents numerical results. Finally, Section VI provides the conclusion and future work.

2. RELATED WORK

Network slicing organizes computing and communication resources of a physical infrastructure to enable flexible support of diverse use case realizations. Each physical network is sliced into multiple virtual networks with network slicing to optimize a specific application. A network slice can have its own network architecture and provision in terms of traffic flow and operation. Network slicing techniques in 5G are expected to simplify the creation and to allow mobile users to handover to another network or request that the service is migrated to a cloud system that is closer to the user's location without an interruption while streaming a video [14]. In order to support the network slicing of a wide variety of wireless networks and enable the communication in such heterogeneous networks, Y-Comm framework has introduced as a generic network architecture. It provides the required functionalities to accommodate different wireless technologies as well as supports the mobile services and advanced handovers [15].

2.1. The Y-Comm framework

The Y-Comm framework is a communication architecture to support vertical handover for multi-nodes mobile users in heterogeneous environments. The architecture has two frameworks. The first one is the peripheral framework which deals with operations on the mobile terminal. The second one is the core framework which deals with functions in the core network to support different peripheral networks. A brief explanation of Y-Comm is given in this paper however, a more detailed explanation can be found in [16]. The Y-Comm reference model is shown in Figure 1.

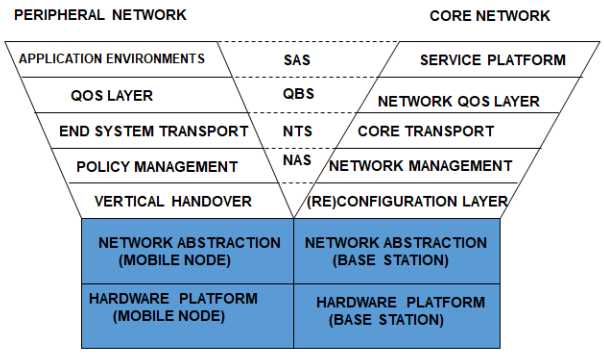

Figure 1. The Y-Comm framework: The reference model [16]

In [6] and [8] a detailed classification of handover and Y-Comm architecture are presented, respectively. However, it is necessary to know how Y-Comm architecture support advance handover in order to perform seamless handover using a mathematical model approach. In Y-Comm, the handover information is managed by the network management layer (NML) in the core framework as shown in Figure 1. The mobile user polls the NML to get information from all available networks as well as the QoS characteristics. The policy management layer (PML) determines where and when a handover should occur based on the obtained information with the direction and speed of the mobile user as well as the QoS. The PML calculates TBVH and the estimated network dwell time (NDT). This information is communicated to the vertical handover layer (VHL) which immediately requests resources to do a handover.

2.2. A Service-oriented framework for mobile services

However, in order to provide a complete set of mechanisms to enable mobile services, it is necessary to develop a new service architecture that allows services to be managed, copied or migrated to support mobile users [15]. This means that services can be managed, copied or migrated to support mobile users. The system provides algorithms that incorporate traffic management as well as the QoS requirements of the flow. This new framework was proposed in [7] as shown in Figure 2 which has six layers are briefly described below:

• The service management layer: This layer manages the service that is being provided. It specifies the functions of the service, registers and obtains a unique service ID. It also specifies the minimum resources required by networking and cloud infrastructure needed to run the service in terms of computing resources, network QoS requirements and storage needs.

• The service subscription layer: This layer handles the functions required for global clients to use the service. It, therefore, allows clients to subscribe to services. It provides the user with a unique client ID, a given service level agreement (SLA) and sets up accounting and payment mechanisms.

• The service delivery layer: The layer is in charge of delivering the service to a given client. It first maps the SLA to a given QoS and ensures that the selected server as well as its networks can meet the required QoS. The service also receives notifications and triggers about handovers and based on these notifications, this layer may replicate or migrate the service closer to the user.

• The service migration layer: The layer handles the replication or migration of services to different cloud platforms to facilitate a good Quality of Experience (QoE) for the mobile user. Migration is done at the request of the service delivery layer.

• The service connection layer: This layer is responsible for managing the connection between a client and the service. It reports changes in transport or network parameters such as bandwidth or delay to the service delivery layer.

• The network abstraction layer: This layer allows the service to interface to different types of networks depending on network architecture and addressing. This layer maps into IP networking with TCP/IP. In more advanced systems such as Y-Comm, this functionality is spread between the QoS and transport layers in core and peripheral frameworks.

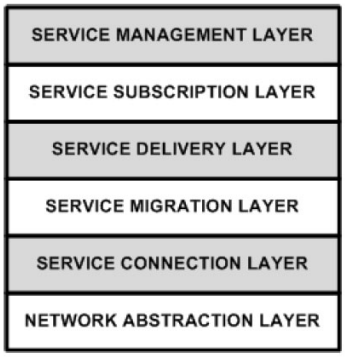

Figure 2. A Service-oriented framework for mobile services

Service migration has been proposed for many environments but has been increased in cloud environments which support virtualization. This is possible due to the virtual machine paradigm which allows entire virtual machines to be migrated [15]. Virtual machine migration can be expensive as the entire virtual machine has to be

migrated. The emergence of container technology, such as Docker, in which containers housing several services can be migrated, is gaining in prominence [1]. Unikernels in which the operating system is bounded and customized to run a single main application is the next emerging specimen in this genre and should make server migration a simpler mechanism from the management viewpoint [1]. However, these efforts assume that the communications architecture does not help facilitate server migration which makes these mechanisms difficult to use in wide area networks (WANs) [15].

The proposed model is a powerful framework that can be used to allow services to migrate from one cloud to another. In order to decide to shift/migrate a service, the works in [6] and [7] have shown how the proposed system should work. In addition, the time taken to shift/migrate the service must be compared with the amount of time the user will be in the region. Thus, the mobility model of the user must be considered. However, in [6] and [7] a simple Markov modeling technique is applied which; there is no generic analytical model as well as a solution approach that can be used for service management. In addition, according to our knowledge, nobody has considered a real vehicular testbed platform to obtain realistic QoS results for such cases. In [17] an analytical modelling approach for large-scale cloud networks is presented from a performance point of view to configure the data center parameters. In addition, analytical models are also presented in [18] for cloud computing. In [19] a hierarchical model is presented for performance analysis of large-scale cloud systems. However, all analytical models presented in the literature have not considered mobile service delivery process with service management framework. In addition, the real-time based scenarios have not been considered. However, in this study all factors and issues mentioned here are considered to get best performance measurements. In this paper, the mobile users decide if a service either migrate (move) to another RSU (network) or shift closer to the mobile user by comparison with network response time obtained based on the slicing factor. The service migration is the network handover between mobile user and RSU while a service moves. In other words, when the mobile users move through different networks, the QoS information of RSUs passes to the mobile users. Thus, the mobile users initiate connections to the new cloud network where another instance of the service is running. In the proposed management framework mobile users have to decide to migrate the service by checking the response time (latency) of the networks (e.g., RSU). If the RSU response time is already high, the mobile user can ask for a cloud network to shift the service closer to the mobile user.

3. THE PROPOSED CLOUD-BASED VEHICULAR NETWORKS AND ANALYTICAL MODEL

In this section performance optimization of cloud-based service management for the vehicular network is investigated together with network slicing concept considering real scenarios and parameters using the Middlesex (MDX) University VANET testbed [23,24]. In order to optimize the use of the network for each mobile device, a queuing model is presented. This is achieved by E2E network slicing when the service localization and migration are desirable. The MDX vehicular testbed has seven RSUs and located as shown in Figure 3 [24]. However, in this paper, examples of different scenarios of three RSUs (namely RSU 1, RSU 2, and RSU 3) are considered for the proposed analytical model to show the effectiveness of the proposed analytical model and solution considering the common scenarios.

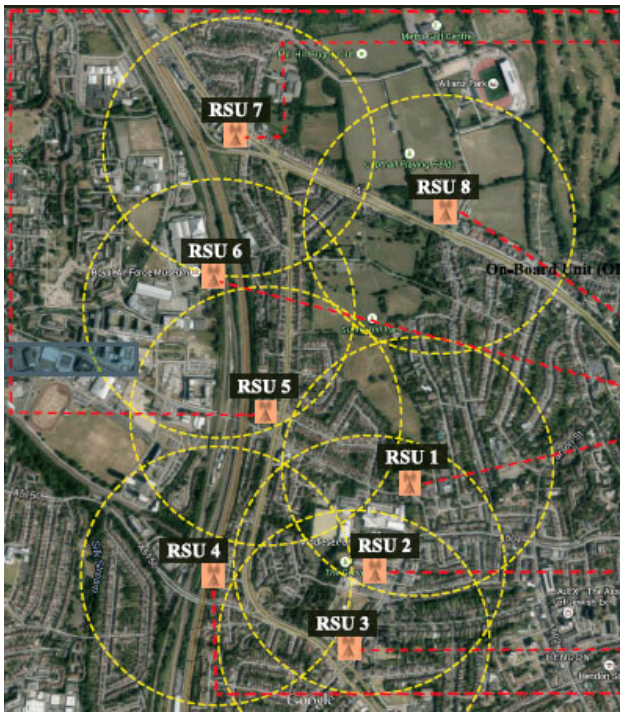

Figure 3. MDX VANET testbed, Hendon, London [23,24]

3.1. The system model

In this section, the proposed system is described as analytical modelling approach to evaluate the QoS of vehicular network considering network slicing using Markov chain. The map of the proposed model considered in this paper is shown in Figure 3. The active mobile user's trajectory is taken from RSU 1 to RSU 3. A mobile user will be connected RSU 1, RSU 2 and RSU 3 during its journey, respectively.

The inputs of the proposed system depend on the mathematical model such as average speed of the users (E[v]), speed of mobile users (v), radius of the RSUs (R), service time (T_S), arrival rates (λ) and network dwell times (T_{dwell}). The requests arrive to each RSU in a Poisson stream at rate λ. The required service time, T_S is independent and is distributed exponentially with rate of μ for each RSU. In addition, network dwell times, T_{dwell} for each RSU are assumed to be exponentially distributed with a mean rate of μ_{dwell} for each network. In other words, $\mu_{dwellR,i}$ and $\mu_{dwellL,i}$ are different handover rates from one RSU to another and are defined as the mobility rates in this paper. An important reactive network slicing parameter is defined and denoted as α. By considering α, the mobile user will get a share of the maximum capacity of the network to which the mobile user is currently attached. In other words, the system will allow the mobile user to calculate the available network resources to use. Furthermore, by using the obtained measurements the mobile users will evaluate other options. The service rate is defined as the perceived rate of service responses that the mobile users are receiving. Hence, the service rates are the factor of α for each RSU. It is assumed that the service rate should be high enough so that it satisfies the request rate from the mobile users. However, the requests queue up by the network and consequently, the response time increases when this condition is not satisfied. Table 1 shows the radius, service and dwell time of all RSUs considered in this paper. Hence, the service rates, as well as the mobility rates can be calculated for proposed analytical modelling considering the real scenario. In addition, the arrival rate λ, the reactive slicing factor of the RSUs α, the average speed of the user (E[v]=50 Mph) are assumed for analytical tractability. In addition, the overlapping distance between RSU 1 & RSU 2 and RSU 2 & RSU 3 is 173m and 828m, respectively.

Table 1. Radius, service and dwell time of the RSUs

Networks	Radius (R)	Service Time (T_S)	Network Dwell Time (T_{dwell})
RSU 1	974m	39.57s	43.57s
RSU 2	1390m	58.19s	62.19s
RSU 3	1140m	47.00s	51.00s

The focus here is the multimedia application such as music, streaming video etc. that user wants to use during its movement. Hence, the challenge here is to use network slicing by handover techniques and server localisation to maintain a reasonable QoS in a real VANET testbed. The different formation of RSUs and mobility along a mobile user path represented as a multi-dimensional model as shown in Figure 4. Figure 4 shows the transitions of the system considered. The transitions of the system are described by *i* and *j*, specifying the RSUs configuration and number of requests in the RSUs, respectively. Thus, the steady-state probabilities of the system are described as $P_{i,j}$s where i and j represent the

number of RSUs and number of requests in the RSUs, respectively. In Figure 4, there are three RSU configurations (i=1,2,3) at the x-axis. *C* is the capacity of each RSU. *C* is the buffer capacity of each RSU. Thus, *C* is taken as 100 requests in the paper.

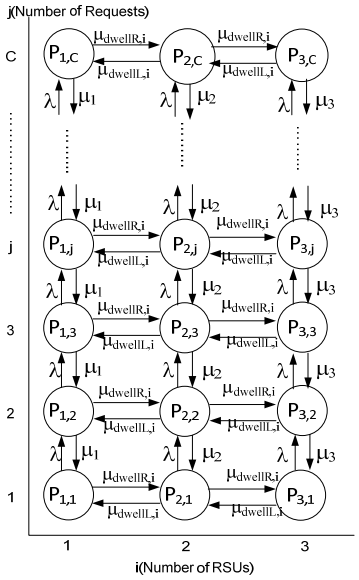

Figure 4. State transit rate diagram of the proposed model

The downward transitions indicate that the requests are being served with service rate μ_i (i=1,2,3) which depends on the number of RSUs. On the other hand, upward transitions take place because of the new request for applications with rate λ to the system. Please note that the proposed system is an analytical model for the mobile user. The proposed model does not analyse the given network which will have many users unlike the previous analytical models [12,17–19,21,22]. Hence, the service rate of the network as seen by the user will vary according to how many other users are using the network. The service rate of networks is defined as the comprehend rate of service responses that the mobile users receive. Thus, the service rates are the factor of α. The lateral transitions indicate mobility scenarios between RSUs. As the mobile users move between RSUs, each column shows mobile user's performance at the given RSU. The mobility of the users between columns is expressed as a mobility rate in the model. $\mu_{dwellR,i}$ and $\mu_{dwellL,i}$, where (i=1,2,3), represent different rates for the mobile user to leave the RSUs by moving to the right-hand side and left-hand side, respectively. In this paper, $\mu_{dwellR,i}$ and $\mu_{dwellL,i}$ rates are obtained from the MDX VANET testbed. Each set of the vertical column represents different PoA to the same RSU. Hence, this gives a fine-grained approach to obtain QoS measurements of a single network from different PoA. In other words, for a realistic scenario, the proposed model takes into account the characteristics of each RSU where each vertical column represents different RSU.

4. MARKOV MODEL OF PROPOSED SYSTEM

In this section, the proposed analytical model is adapted and generalized for such systems. The solution of the proposed model can be used for any type of network and configuration for performance analysis.

4.1. Solution of the proposed system

In this study, state probabilities, $P_{i,j}$s are obtained by using the solution of the successive over relaxation (SOR) method. Thus, the well-known system of balance equations is obtained in order to solve the system. Figure 4 indicates the state transit rate diagram of the proposed system however, when the system generalized we can increase the number of RSUs and request as we like. For instance, N is the maximum number of RSUs and *C* is the maximum number of requests that can be allocated in the system. MATLAB package is used for the solution of the proposed model. The system state probabilities are obtained with linear equations of the form of *Ax=B* as shown in equation 1 when the balance equations are derived properly.

$$\begin{bmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N*C} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N*C} \\ \vdots & \vdots & \vdots & \vdots \\ A_{N*C,0} & A_{1,N*C} & \cdots & A_{N*C,N*C} \end{bmatrix} * \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N*C} \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{N*C} \end{bmatrix} \quad (1)$$

In the proposed system, A is of size N*C. x is a column vector of unknown state probabilities (P_i) where i=0,1,...,N*C. B consists of the scalars in the balance equations. A is symmetric, real and positive definite matrix. Cholesky factorization is well-known direct method for solving such linear system. It can be computed by form of Gaussian elimination. Hence, the Cholesky

method decomposes matrix A in the form of $A=L^TL$ where L is upper triangular matrix and L^T its transpose. L can be chosen so that its diagonal elements are strictly positive and it is unique. Then, a substitution performed in the equation of Ax=B. Hence, the resulting equation is $L^TLx=B$ which can be solved for x. Let Lx=Y and the equation can be written as $L^TY=B$. Therefore, first solve $L^TY=B$ for Y and then solve Lx=Y for x which will give the final solution. If the Cholesky factorization fails, an asymmetric, indefinite factorization is performed. Further details can be found in [25] on the Cholesky factoring. B_i vector is denoted, where $B=\{0,0,...,0,1\}$. Then, the solution of the system is as follows: $A_i.P_i=B_i$ where i=0,1,....,N*C. When the model is in steady state, transition rates used to obtain the matrix A considering the balance equations are in an equilibrium.

i=0 and j=0;

$$P_{i,j} = \frac{\mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \mu_i P_{i,j+1}}{\lambda + \mu_{\text{dwellR,i}}} \tag{2}$$

i=0 and 0<j<C;

$$P_{i,j} = \frac{\mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \mu_i P_{i,j+1} + \lambda P_{i-1,j}}{\lambda + \mu_{\text{dwellR,i}} + \mu_i} \tag{3}$$

i=0 and j=C;

$$P_{i,j} = \frac{\mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \lambda P_{i-1,j}}{\mu_{\text{dwellR,i}} + \mu_i} \tag{4}$$

0≤i<N and j=0;

$$P_{i,j} = \frac{\mu_i P_{i,j+1} + \mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \mu_{\text{dwellR,i-1}} P_{i-1,j}}{\lambda + \mu_{\text{dwellR,i}} + \mu_{\text{dwellL,i}}} \tag{5}$$

***0≤i<N and* 1≤j<C;**

$$P_{i,j} = \frac{\mu_i P_{i,j+1} + \mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \lambda P_{i,j-1} + \mu_{\text{dwellR,i-1}} P_{i-1,j}}{\lambda + \mu_{\text{dwellR,i}} + \mu_i + \mu_{\text{dwellL,i}}} \tag{6}$$

***0≤i<N and* j=C;**

$$P_{i,j} = \frac{\mu_{\text{dwellL,i+1}}\, P_{i+1,j} + \lambda P_{i,j-1} + \mu_{\text{dwellR,i-1}} P_{i-1,j}}{\mu_{\text{dwellR,i}} + \mu_i + \mu_{\text{dwellL,i}}} \tag{7}$$

i=N and j=0;

$$P_{i,j} = \frac{\mu_N P_{i,j+1} + \mu_{\text{dwellR,N-1}} P_{N-1,j}}{\lambda + \mu_{\text{dwellL,i}}} \tag{8}$$

i=N and 1≤j<C;

$$P_{i,j} = \frac{\mu_N P_{N,j+1} + \lambda P_{N,j-1} + \mu_{\text{dwellR,N-1}} P_{N-1,j}}{\lambda + \mu_N + \mu_{\text{dwellL,i}}} \tag{9}$$

i=N and j=C;

$$P_{i,j} = \frac{\lambda P_{N,L-1} + \mu_{\text{dwellR,N-1}} P_{N-1,C}}{\mu_N + \mu_{\text{dwellL,N}}} \tag{10}$$

When all the steady-state probabilities, $P_{i,j}$s, are obtained, various QoS measurements could be easily computed. Thus, the mean queue length (MQL_i), throughput (γi) and mean response time (MRT_i) of all RSUs are computed using equations 11, 12, and 13, respectively.

$$MQL_i = \sum_{i=0}^{N} i \sum_{j=0}^{C} P_{i,j} \tag{11}$$

$$\gamma_i = \sum_{i=0}^{N_i} \mu_i \sum_{j=0}^{C} P_{i,j} \tag{12}$$

$$MRT_i = \frac{MQL_i}{\gamma_i} \tag{13}$$

The proposed network slicing framework heavily depends on measuring the mean response time and mean queue length of RSUs. Hence, the decision of RSU selection can be done by the mobile users obtaining the performance measurements. The proposed service management evaluates the various QoS measurements of RSUs and prepares a list of available as well as suitable networks that how much of the network slice of mobile users need.

5. SCENARIO BASED APPLICATION AND NUMERICAL RESULTS

In this section, numerical results are presented for the cloud-based vehicular network for the mobile service delivery framework considering network slicing by calculating mean response time and mean queue length of RSUs when a user is mobile while streaming a video.

In this paper, the wireless footprint method is also assumed in order to obtain mobility pattern information for the mobile user as described in [6]. The user mobility pattern information can be provided based on a user's past mobility patterns. Thus, this method can be used to estimate the next network that the user may join to obtain best QoS for the mobile users. The requirements and modelling issues are considered for video application. In addition, in terms of the QoS and the time the user spends in these networks may vary because of different RSU sizes due to the heterogeneous environment considered. However, the proposed model can be adapted

easily to different types of application as well as the traffic based on the network specifications. Hence, in order to stream a video without interruption in a mobile environment, the analysis done in [20] is followed to obtain optimal latencies for the service decision making process. In [20] authors show that if delay, D < 40ms, the mobile user can carry on using the same RSU without interruption. If 40ms ≤ D ≤ 80ms, the mobile user should handover to better RSU in order to stream video. On the other hand, if 80ms < D, the mobile user can appeal to switch the service closer to the user to maintain QoS. Otherwise, the mobile user will lose the connection.

5.1. Real time scenarios of video streaming

This section includes two different scenarios of a mobile user may be commonly encountered by streaming a video in the real world. The results of these case studies can assist in understanding the proposed model and multimedia content quality adaptation. Furthermore, these results prove analytically that the model is functioning as expected. The system parameters used are mainly taken from MDX VANET testbed as well as the relevant literature [6,7,17–24]. To demonstrate how this model behaves in different scenarios, we study a case where the RSUs have different service rates and keep changing according to the intensity of mobile user in the system. α is the network slicing factor. In other words, it is the probability of service rate changes in all RSUs. Hence, $\mu_i\alpha$ gives service rate changes in all RSUs while streaming a video. All scenarios express a case where a mobile user may be handed-over to any available RSUs or requested from the RSUs to shift the service closer to the mobile user in order to stream a video without interruption. In the considered scenarios, the activity of all these networks is very much dependent on what is happening during the day.

5.1.1. Scenario I

In this scenario, we consider a mobile user moving from RSU 1 to RSU 3. The RSU 1 and RSU 3 have slightly equal service rates as μ_1=0.0253 reqs/sec and μ_3=0.0212 reqs/sec, respectively. The RSU 2 is located between RSU 1 and RSU 3. The service rate of RSU 2 is μ_2=0.017 reqs/sec as given in Table 2. The RSUs based request rate, λ is 0.01 reqs/sec. The results will let us to predict service decision making process across various RSUs through a user's path.

Table 2. Service and mobility rates of RSUs for scenario I

Networks	Service Rate (reqs/secs)	$\mu_{dwellR,i}$ (reqs/secs)	$\mu_{dwellL,i}$ (reqs/secs)
RSU 1	0.0253	0.023	0.023
RSU 2	0.017	0.016	0.016
RSU 3	0.0212	0.019	0.019

As we see from the results in Figures 5 and 6, the mobile user requests are queued upon the RSU 2 which has an insufficient service rate when compared to other networks. This leads to increase in mean response time and mean queue length accordingly. Consequently, there will be performance degradation as the network will be busy if the mobile user joins the RSU 2. On the other hand, RSU 1 had better performance results since the reactive service rate ($\mu_i\alpha$) is higher than the other RSUs.

Finally, the RSU 2 had higher latency and therefore, the performance is insufficient for the mobile user.

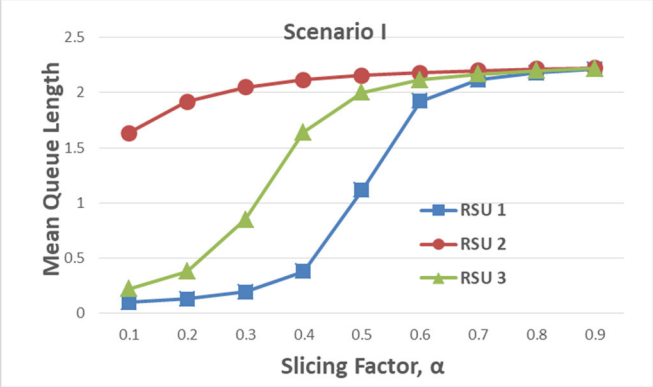

Figure 5. Mean queue length results as a function of slicing factor for scenario I

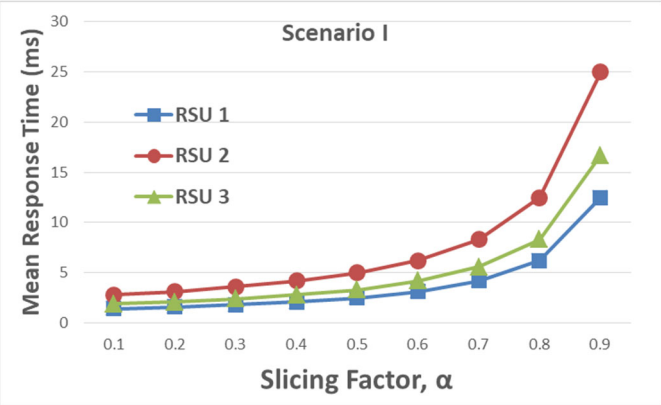

Figure 6. Mean response time results as a function of slicing factor for scenario I

In addition, when the system is significantly busy α=0.9, the mobile user can be requested from other RSUs to migrate their services to closer. For example, it is assumed that a mobile user is out of the coverage area of RSU 1 and RSU 3. It is connected to RSU 2 and also it moves out of the coverage area of RSU 2 towards RSU 3. Figures 5 and 6 show the results from this case and we see that the performance of RSU 2 is insufficient to stream a video. The RSU 3 is also not capable of maintaining a high QoS after some threshold point (e.g. $\alpha > 0.8$). In order to achieve efficient utilisation of resources, the mobile user can request to RSU 3 to shift the service closer to it and temporarily stream music from there or shift the service to the RSU 1 without joining to RSU 3 and stream video from there.

5.1.2. Scenario II

In this case, we consider a heavy traffic in the system based on the service rates provided by the networks and the mobility rates of the user as shown in Table 3.

Table 3. Service and mobility rates of RSUs for scenario II

Networks	Service Rate (req/secs)	$\mu_{dwellR,i}$ (req/secs)	$\mu_{dwellL,i}$ (req/secs)
RSU 1	0.067	0.023	0.023
RSU 2	0.034	0.016	0.016
RSU 3	0.05	0.019	0.019

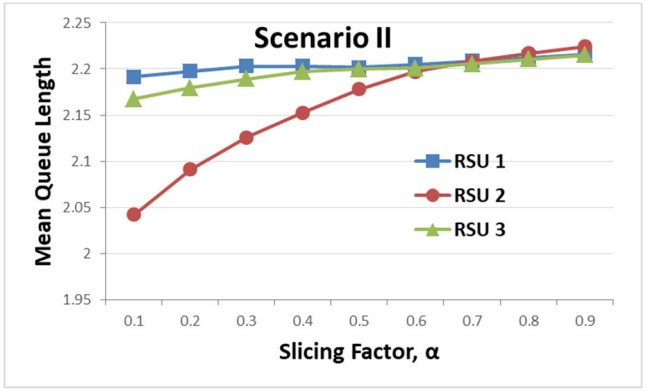

Figure 7. Mean queue length results as a function of slicing factor for scenario II

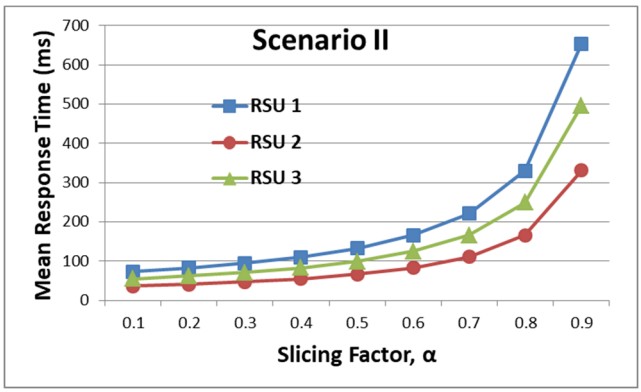

Figure 8. Mean response time results as a function of slicing factor for scenario II

Figures 7 and 8 also show the mean queue length and the mean response time results as a function of α for scenario II, respectively. As clearly seen from the figures, the performance results of the RSU 2 are better than the performance results of the RSU 1 and RSU 3. Thus, the mobile user can be handed-over to RSU 3 even though the serving capability of RSU 3 is not sufficient to stream a video without interruption when $0.1<\alpha<0.3$. However, a mobile user should find an alternative network to be handed-over in order to stream a video without interruption after this. In addition, the mobile user has to request a service migration from the available RSUs in order to stream a video when $\alpha \geq 0.4$ for the RSU 3. Table 4 shows the decision making results to find the sequence of RSUs that the mobile users will attach. It will either handover to available networks or request to service localization.

Table 4. Mean resposne times of RSUs for scenario II

α	RSU 1	RSU 2	RSU 3
0.1	73.9s	37s	55.5s
0.2	83.1s	41.6s	62.4s
0.3	95s	47.6s	71.4s
0.4	110.7s	55.5s	83.2s
0.5	132.8s	66.6s	99.9s
0.6	165.9s	83.2s	124.8s
0.7	220.8s	110.9s	166.3s
0.8	330.1s	166.3s	249.1s
0.9	653.8s	331.7s	496.3s

Hence, for such cases, we may consider the mobile user to start streaming the video from the RSU 1 as it firstly hits the coverage area of the RSU 1. When the mobile user reaches the coverage area of RSU 2, after some time, the mobile user might have to handover the service to available RSUs when $0.1<\alpha<0.6$ as shown in Table 4. However, the mobile user can be handed-over to the other RSUs to carry on its service without interruption when $\alpha\leq0.5$ for RSU 2.

Table 5. Comparison of mean response time results in the absence of the proposed approach (PM: Proposed model and APM: Absence of the proposed model)

	RSU 1		RSU 2		RSU 3	
α	PM	APM	PM	APM	PM	APM
0.1	73.9s	76.4s	37s	38s	55.5s	56.8s
0.2	81.1s	90.2s	41.6s	43s	62.4s	63.3s
0.3	95s	98.7s	47.6s	50.6s	71.4s	75.2s
0.4	110.7s	120.8s	55.5s	60.2s	83.2s	89.7s
0.5	132.8s	143.8s	66.6s	80.8s	99.9s	110.9s
0.6	165.9s	200.1s	83.2s	90.3s	124.8s	140.1s
0.7	220.8s	300.6s	110.9s	150.4s	166.3s	183.2s
0.8	330.1s	440.2s	166.3s	223.5s	249.1s	320.4s
0.9	653.8s	700.2s	331.7s	452.7s	496.3s	520.8s

In order to show the effectiveness of the proposed model the performance results presented in Table 4 are also considered and presented in Table 5 in the absence of the proposed approach. It is clearly evident from the Table 5 that the proposed model increases the system performance significantly. The mean response time results show that all of the RSUs become quickly busy when the proposed model is not considered. However, the response times of each RSUs decreases (in other words, each RSUs respond much quicker) when the proposed model is applied. Thus, the QoS of each RSUs increase in terms of response time when the proposed model is considered.

Both scenarios demonstrate how various cases may be constructed considering the proposed model. It is clearly seen from the results that network slicing factor α, affect the system performance significantly. Hence, the proposed model can be used to analyze the overall performance of each scenario. This will lead the mobile users to choose an existing network to be served without interruption. In other words, the mobile users will decide to select the available RSU with better performance. Thus, this model can be applied as a network slicing framework in vehicular environments to improve QoS.

6. CONCLUSION

This paper proposed an analytical modelling approach and QoS provisioning for E2E network slicing service management framework in vehicular environments. The proposed model requests services directly from the available RSUs rather than offered services by the physical resources. Thus, this will simplify the process for end users and also gives opportunities to direct mobile user requests to the most appropriate RSUs or to run a service closer to the mobile user location. The proposed model calculates the increase in delay as a user moves while streaming a video. It offers the perspective of considering the response time of the each RSU in the system before the user requests a service. The mobile user can make a decision by analysing the RSU. It may either handover the service to available RSUs or requests to migrate the service to closer

location without an interruption while streaming a video. Overall, the results were convincing and show that the proposed approach can be useful in real networks to maintain QoS for the mobile user in vehicular networks.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gelnford Mapp, Dr. Fragliskos and Mr. Vishnu V. Paranthaman for their guidance and helpful suggestions. Dr. Glenford Mapp has a funding from Department of Transport (DfT) and he is the head of Middlesex University VANET research team. Dr. Yönal Kırsal is a visiting fellow (international collaborator) of the MDX VANET research team.

REFERENCES

[1] Q. Li, G. Wu, A. Papathanassiou, L. Wei, “End-to end network slicing in 5G wireless communication systems", *In Proc. of ESTI workshop on future radio technologies: Air interfaces*, Jan. 27-28, 2016.

[2] P. K. Agyapong, M. I., "Design Considerations for a 5G Network Architecture", IEEE Communications Magazine, vol. 52, no. 11, 2014, pp. 65-75.

[3] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility Resource Management and Challenges", *Communications Magazine IEEE*, vol.55, 2017, pp. 138-145.

[4] Ericsson white paper, 5G systems, http://www.ericsson.com/res/docs/whitepapers/whatis-a-5g-system.pdf, Jan. 2015. Accessed: 2015-05-29.

[5] Nokia, Dynamic end-to-end network slicing for 5G, White paper, Finland, 2016

[6] F. Sardis, G. Mapp, J. Loo, M. Aiash and A. Vinel, “On the Investigation of Cloud-Based Mobile Media Environments with Service-Populating and QoS-Aware Mechanisms," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 769-777, June 2013.

[7] F. Sardis, “Exploring Traffic and QoS Management mechanisms to support mobile cloud computing using service localization in heterogeneous environments", School of Science and Technology, Middlesex University, August 2014, PhD Thesis.

[8] G. Mapp, F. Katsriku, M. Aiash, N. Chinnam, R. Lopes, E. Moreira, R. P. Vanni, and M. Augusto, “Exploiting Location and Contextual Information to Develop a Comprehensive Framework for Proactive Handover in Heterogeneous Environments" *Journal of Computer Networks and Communications*, pp. 1-17, 2012.

[9] G. Mapp, F. Sardis and J. Crowcroft, “Developing an implementation framework for the Future Internet using the Y-Comm architecture SDN and NFV", *IEEE NetSoft Conference and Workshops (NetSoft)*, Seoul, pp. 43-47, 2016.

[10] X., Dionysis, “Handover decision for small cells: Algorithms, lessons learned and simulation study", *Computer Networks*, 100, 2016, pp. 64-74.

[11] F. Shaikh, Intelligent proactive handover and QoS management using TBVH in heterogeneous networks [Ph.D. thesis], School of Engineering and Information Sciences, Middlesex University, January 2010.

[12] Y. Kirsal, “Analytical Modelling of a New Handover Algorithm for Improve Allocation of Resources in Highly Mobile Environments", *International Journal of Computers Communications and Control*, 2016, 11, 6, pp. 755-770.

[13] D. Cottingham, I. Wassell, and R. Harle, “Performance of IEEE 802.11a in vehicular contexts", *In Proceedings of IEEE 65th Vehicular Technology Conference*, 2007, pp. 854-858.

[14] NGMN Alliance, “5G White Paper”, February 2015.

[15] A. Ghosh, V. V. Paranthaman, G. Mapp, O. Gemikonakli, J. Loo, "Enabling seamless V2I communications: toward developing cooperative automotive applications in VANET systems", IEEE Commun. Mag., vol. 53, no. 12, pp. 80-86, 2015.

[16] G. Mapp, F. Shaik, D. Cottingham, J. Crowcroft, J. Baliosian, "Y-Comm: a global architecture for heterogeneous networking", In Proceedings of the 3rd international conference on Wireless internet (WICON '07). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, Article 22, 5 pages.

[17] D. Bruneo, "A stochastic model to investigate data center performance and QoS in iaas cloud computing systems", *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, 2014, pp. 560-569.

[18] J. Vilaplana, F. Solsona, I. Teixid, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing", *The Journal of Supercomputing*, vol. 69, no. 1, 2014, pp. 492–507, 2014.

[19] R. Ghosh, F. Longo, V. K. Naik, and K. S. Trivedi, "Modeling and performance analysis of large scale iaas clouds", *Future Generation Computer Systems*, vol. 29, no. 5, 2013, pp. 1216-1234.

[20] P. Vidales, J. Baliosian, J. Serrat, G. Mapp, F. Stajano and A. Hopper, "Autonomic system for mobility support in 4G networks", *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 12, pp. 2288-2304, Dec. 2005.

[21] Y. Kirsal, E. Ever, A. Kocyigit, O. Gemikonakli, G. Mapp, "Modeling and analysis of vertical handover in highly mobile environments", *The Journal of Supercomputing*, 71, pp. 4352-4380, 2015.

[22] R. Guillaume, A. G. Andres, A. Ben, "LTE Advanced and Next Generation Wireless Networks Channel Modeling and Propagation", John Wiley and Sons Ltd. 2013.

[23] Building an Intelligent Transport Information Platform for Smart Cities, Report, http://www.vanet.mdx.ac.uk/, 2016. Accessed: 2017-09-29.

[24] Building a Connected Vehicle Testbed to study the development and deployment of C-ITS in the UK, http://www.its-ukreview.org/building-a-connected-vehicle-testbed-to-study-the-development-and-deployment-of-c-its-in-the-uk/,2016, Accessed : 2017-09-30.

[25] N. J. Higham Functions of Matrices: Theory and Computation. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2008, ISBN 978-0- 898716-46-7.