Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher’s definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher’s website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. Apr. 2020
Network slicing management & prioritization in 5G mobile systems

Menglan Jiang, Massimo Condoluci, Toktam Mahmoodi
King’s College London, London, UK
{menglan.jiang, massimo.condoluci, toktam.mahmoodi}@kcl.ac.uk

Abstract—5G mobile network is expected to serve flexible requirements hence dynamically allocate network resources according to the demands. Network slicing, where network resources are packaged and assigned in an isolated manner to set of users according to their specific requirements, is considered as a key paradigm to fulfill diversity of requirements. There will clearly be conflicting demands in allocation of such slices, and the effective provisioning of network slicing poses several challenges. Indeed, network slicing has a twofold impact in terms of user/traffic prioritization as it dictates for the simultaneous management of the priority among different slices (i.e., inter-slice) and the priority among the users belonging to the same slice (i.e., intra-slice). In this paper, we propose a novel heuristic-based admission control mechanism able to dynamically allocate network resources to different slices in order to maximize the satisfaction of the users while guaranteeing to meet the requirements of the slices they belong to. Through simulations, we demonstrate how our proposal provides (i) higher user experience in individual slices, (ii) increased utilization of network resources and (iii) higher scalability when the number of users in each slice increases.

Index Terms—5G; slicing; admission control; QoS; heuristic.

I. INTRODUCTION

With the fast growth of wireless network technologies (e.g. 5G) and ever-increasing demand for services with high Quality of Service (QoS) demand [1], the management of network resources becomes an always more challenging task that needs to be properly designed in order to improve network performance. In this scenario, network slicing [2] is gaining an always increasing importance as an effective way to introduce flexibility in the management of network resources. A slice is a collection of network resources, selected in order to satisfy the requirements (e.g., in terms of QoS) of the service(s) of the service(s) to be provided by the slice [3]. The intention of slicing is to introduce flexibility and higher utilization of network resources by providing only the network resources necessary to fulfill the requirements of the slices enabled in the system.

An enabling aspect of network slicing is the virtualization of network resources, which allows operators to share the same physical resources in a flexible, dynamic manner in order to exploit the available resources in a more efficient way [4]. Virtualization of network resources is currently investigated in literature especially by focusing on the virtualization of network functionalities [4], [5], [6], [7]. Due to the diverse QoS requirements and the limitation of network resources, efficiently allocate network resources among service slices and UEs is a significant issue [8]. In this field, further study is needed on the virtualization of radio resources in order to perform the admission control and the resource allocation for network slices. Indeed, an important aspect to be considered is the way radio resources are allocated to different slices in order to meet the requirements of such slices. The task relevant to radio resource allocation becomes more challenging with network slicing, as it introduces a two-tier priority in the system. The first tier refers to the priority of different slices, i.e., inter-slice priority, as each slice has its own priority defined according to the agreements between the slice owner and the network provider. The second tier refers to the priority among the users of the same slice, i.e., intra-slice priority.

When looking at the solutions exploited over current 4G systems to manage radio resources, it clearly emerges that 4G networks are able to maximize the QoS of the served users but are not able to perform the resource allocation in slicing environments [9]. This limitation is due to the fact that resource allocation in 4G systems is performed by associating a priority to the service requested by the user equipment (UE). This approach thus fails when considering that in 5G systems different UEs may belong to different slices with different priorities, and thus such UEs should be managed by considering the priority of the slice they belong to plus the priority of the service they require.

In this paper, we propose a novel heuristic-based admission control mechanism. As shown in Fig. 1, the proposed admission control mechanism exploits a two-tier priority levels. Our proposal is based on the idea that network slices communicates to an admission control entity the desired QoS level. The admission control mechanism, based on the priority of the slice, decides about serving the slice. Finally, according to the inter- and intra-slice priority, the virtual network allocates the physical radio resources to the UEs of the admitted slices. According to the decision of the admission control, the resource allocation task is performed with the aim to maximize the quality of experience (QoE) of the users within each slice, by considering the inter-slice priority. In this paper, the QoE is measured by considering the effective throughput experienced by the users, normalized according to their maximum requested data rate. With this aim, the resources allocated to a slice with low priority could be reduced, if necessary, down to the minimum amount able to meet the basic QoS requirements in order to admit new slice(s) with higher priority. So doing, our proposal dynamically changes the amount of network resources allocated to network slices...
according to the traffic load without affecting the QoE of the users and while improving the network utilization. To summarize, the main contributions of this paper can be stated as follows:

- A novel heuristic based admission control mechanism with two-tier priority level has been proposed in our virtualized 5G system model. The proposed admission control mechanism dynamically set the resources allocated to enabled slices according to the current traffic load.
- Inter-slice and intra-slice priority order has been taken into account for designing the QoE maximization problem of resource allocation task. Considering priority orders for QoE function can improve the satisfactory level of UEs and network utilization.

The remainder of this paper is organized as follows. Section II reviews the state of art on virtual resource allocation in different kinds of network technologies. After elaborating our system model in Section III, our proposed admission control mechanism will be described in Section IV. Section V presents simulation study and performance observations. Finally, summary of our work is given in Section VI.

II. RELATED WORKS

In literature, several solutions for efficiently supporting virtualization of network resources have been designed to improve the QoE of UEs and network resource utilization [6].

An efficient wireless network virtualization for Long Term Evolution (LTE) systems has been proposed in [10], which proposes a slicing scheme to efficiently allocate physical resource blocks to different service providers (SPs) in order to maximize the utilization of resources. The scheme is dynamic and flexible for addressing arbitrary fairness requirements of different SPs. Similarly, [11] proposed a framework for wireless resource virtualization in LTE system to allow sharing of radio resources between mobile network operators. An iterative algorithm has been proposed to solve the Binary Integer Programming (BIP) with less computational overhead. Nevertheless, above considered schemes do not consider the priority among different slices as well as the priority among the users within the same slice.

For the limitation of network resources, the admission control mechanism can be implemented to improve communication reliability and network utilization. In [12], a joint resource provisioning and admission control mechanism has been proposed aiming to maximize the total rate of virtualized networks based on their channel state information. An iterative slice provisioning algorithm has been proposed to adjust minimum slice requirements based on channel state information but without considering global resource utilization of the network as well as inter- and intra-slice priority.

In [13], a mechanism for allocating downlink network resources has been proposed. The mechanism decides to accept a novel service only if the provisioning of this new service does not affect the throughput of the services in the cell. As a consequence, this work does not take into consideration the dynamic modification of the QoE experienced by mobile users in order to increase network capacity and resource utilization.

Centralized joint power and admission control mechanism for prioritized multi-tier cellular networks has been proposed in [14]. The mechanism has been developed to admit users with higher priority level in order to maximize the number of users. In this case, the priority is only considered at the user level and, thus, this work fails in guaranteeing differentiation in case users belong to slices with different priorities.

III. OUR SYSTEM MODEL

As depicted in Fig. 2, our model consists of four main elements: the service slice layer, the virtual network layer, the physical resources, and the admission control manager. The first three elements will be explained in the remainder of this Section, while the admission control manager will be treated in Sec. IV.
A. Service Slices

The service slices present different services (e.g., car management, TV streaming and web browsing) which require resources to be served. We indicate with $S = \{1, 2, 3, ..., S\}$ the set of slices in the virtual network. Each slice $s$ has a set of UEs, such a set is denoted by $U_s = \{1, 2, ..., U_s\}$. Each slice $s$ performs a request to the admission control in terms of QoS constraints. In this paper, we model such a request with $R_s^{\min}$ and $R_s^{\max}$, which denote the minimum and maximum data rates associated to the slice $s$, respectively.

Each slice $s$ is characterized by a priority, $\rho_s$, where such priorities are defined with the constraint that $\sum_{s \in S} \rho_s = 1$. Similarly, each user $u$ belonging to the slice $s$, i.e., $u_s$, is characterized by a priority $\mu_{us}$, where $\sum_{u \in U_s} \mu_{us} = 1$.

B. Virtual Network

The virtual network layer provides an abstraction of the physical network resources. According to the decisions of the admission control, the virtual network slices the resources of network to accommodate different slices. The virtual network receives the requests of different slices in terms of UEs to be served for each slice, and performs the subsequent allocation of physical resources according to the inter- and intra-slice priority while taking into account the QoE of UEs.

With this aim, (10), we can define:

$$q_{us} = \left( \frac{r_{us}}{R_s^{\max}} \right)$$

(1)

as the QoE of UE $u$ in the slice $s$; $r_{us}$ is the data rate of the UE $u$ in the slice $s$. The overall QoE of users belonging to slice $s$ can be computed as:

$$q_s = \sum_{u \in U_s} (q_{us})^{\mu_{us}}$$

(2)

Finally, we can define:

$$Q = \sum_{s \in S} (q_s)_{\rho_s}$$

(3)

as the overall QoE experienced by all the UEs of all slices.

The virtual network assigns the resources on a scheduling-frame basis. We define with $q_{us}^t$, $q_s^t$ and $Q^t$ the QoE in a generic scheduling frame $t$. Accordingly, we can also define the time-average QoE values as follows:

$$E[q_{us}] = \frac{1}{T} q_{us}^t$$

(4)

$$E[q_s] = \frac{1}{T} q_s^t$$

(5)

$$E[Q] = \frac{1}{T} Q^t$$

(6)

where $T$ is the overall number of considered scheduling frames.

C. Physical resources

The physical resources refer to the radio resources available in the virtual network. For the sake of simplicity, we refer to the downlink channel of one macro-cell. The total available bandwidth is denoted by $B$ MHz. The set $M = \{1, 2, ..., M\}$ represents the available sub-channels, where the bandwidth of the generic sub-channel $m$ is $b_m = \frac{B}{M}$. The total transmit power $P_{TOT}$ is uniformly allocated to each sub-channel, i.e., $p_m = \frac{P}{M}$.

When assigning the physical resources, we consider the channel conditions of the UEs. We assume that channel condition is determined by transmission path loss and shadowing components [15]. The path loss is defined in Table I and the shadowing fading path loss is assumed to be a Gaussian random variable with zero mean and $\sigma$ standard deviation equal to $8dB$ [15]. Therefore, the path loss is based on the distance value $d_{us}$ between a generic UE and the macro-cell, which is given by Equation 7.

$$PL(d_{us}) = 128.1 + 37.6 \log_{10}(d_{us}) + \log_{10}(X_{us})$$

(7)

where $X_{us}$ is the log-normal shadow fading path loss of the UE [15].

We also assume that the macro-cell receives perfect channel gain information from all UEs belong to different service slices, where $h_{m,us}$ is the sub-channel gain for the UE $u$ within slice $s$ and can be defined as $h_{m,us} = 10^{-PL(d_{us})/10}$ [15]. The data rate of the UE with slice $s$, denoted with $r_{us}$, can be described in Equation (8) [10].

$$r_{us} = \sum_{m \in M} \alpha_{m,us} b_m (1 + \frac{p_m |h_{m,us}|^2}{N_0 b_m})$$

(8)

where $N_0$ is the noise spectral density and $\alpha_{m,us}$ is the situation of the UE $us$ which has been defined as Equation (9).

$$\alpha_{m,us} = \begin{cases} 1 & \text{if sub-channel } m \text{ is assigned to } u_s \\ 0 & \text{otherwise} \end{cases}$$

(9)

IV. TWO-TIER ADMISSION CONTROL AND RESOURCE ALLOCATION

In this section, we describe our proposed approach for two-tier admission control and resource allocation.

A. Admission Control Strategy

An heuristic-based prioritized admission control mechanism has been designed in Algorithm 1. This mechanism can be used to deal with the arrivals of new slices or users and provides a global optimization of the resources allocated to service slices. For the sake of simplicity, Algorithm 1 refers to the admission control of novel UEs belonging to the same slice. The steps of our proposed admission control mechanism can be used for admission control of new slices, by easily adapting the parameters under consideration. When the new UE enters the network, by considering the QoE of the users in the same slice, we can derive an acceptance probability of the novel user in the virtual network by considering the
constraints in terms of intra-slice priority as well as the QoE of served UEs. In our admission control, a new UEs is accepted if the available resources are sufficient to guarantee to satisfy at least the requirement on the minimum data rate. The set of accepted users are thus provided as input to the resource allocation procedure.

Algorithm 1: Heuristic based Admission Control Algorithm of New Users

for $t := 1$ to $T$ do
  for $s := 1$ to $S$ do
    for $u := 1$ to $U$ do
      for $m := 1$ to $M$ do
        Calculate $q_u \forall u \in U_s$;
        find UE $x_s$ with the max QoE;
        find UE $j_s$ with the max QoE;
        while a new UE $u'_s \in U_s$ enters the network do
          Calculate the new QoE value of $u'_s$: $q_{u'_s}$;
          then, find the neighbor QoE value of $u'_s$: $q_{u'_s}$;
          if $q_{u'_s} - q_{u_s} > 0$ then
            if $E[q_{u'_s}] < q_{u'_s}$ then
              Inject UE $u'_s$;
              check priority order;
              if the priority order are the same then
                $x_s$ will be replaced by the new UE; else
                $j_s$ will be replaced by the new UE;
              end
            end
            else
              Do not admit UE $u'_s$;
            end
            else
              generate accept probability $p = \frac{\alpha_{m,u} b_m}{\rho_{s}}$;
              then, the new UE will be rejected based on the probability $p$;
            end
          end
        end
      end
    end
  end
end

B. Resource Allocation

The overall problem under consideration during the resource allocation step is the maximization of the QoE of UEs, by simultaneously considering the inter- and intra-slice priority. This problem can be formulated as in Equation 10.

**P1:**

$$\max_{s \in S, u \in U_s} \left( \frac{r_{u_s}}{R^{\text{max}}_{u_s}} \right)^{\rho_s}$$

subject to,

$$\sum_{m \in M} \sum_{s \in S} \alpha_{m,u} b_m \leq B,$$ (11a)

$$R^{\min}_{u_s} \leq r_{u_s} \leq R^{\max}_{u_s},$$ (11b)

where, constraint (11a) indicates that the amount of allocated sub-channels cannot overcome the maximum available bandwidth; this constraint implicitly refers to the orthogonality of assigned resources, too. Constraint (11b) indicates that the received data rate by UE $u_s$ is restricted by the requirements of the associated slice $s$. It is worth noting that, in Equation (10), the QoE is a number lower or equal than 1; as a consequence, the higher the priority of a slice, the lower the value of $\rho_s$. This happens similarly for the users, i.e., the higher the priority of a user, the lower is the value $\mu_{u_s}$.

The resource allocation procedure is performed by considering the physical resources available in the network as well as the channel conditions of the UEs.

V. PERFORMANCE INVESTIGATION

This Section provides an performance comparison of our proposal with a legacy 4G resource allocation algorithm. With this aim, we implemented a resource allocation algorithm where network resources are allocated in order to maximize the overall QoE of users by taking into account their QoS requirements (minimum and maximum data rate) as well as the priority of each user. As a consequence, our considered benchmark, hereinafter named 4G service allocation (4G-SA), is a single-tier priority algorithm which does not take into consideration the possibility that UEs belong to different slices. The differentiation among the UEs refers only to a different requested service. For the sake of completeness, we consider two different schemes. The first one, which is referred to as 5G Slice allocation (i.e., 5G-SA), implements the the resource allocation scheme in Sec. IV-B. The second solution, hereinafter 5G-AC-SA, takes into consideration the admission control procedure in Sec. IV-A, which is performed as a first step before the resource allocation. The reason behind this choice is to highlight the impact of the admission control in the management of network resources.

In our simulations, we consider that the arrival rate of UEs is uniformly distributed within the whole simulation period. The overall number of UEs is uniformly distributed among the considered slices. The priority of UEs within the same slice is randomly generated with the constraint of having a sum equal to 1. In case of 4G-SA, the priority of UEs are the same of those considered for 5G-SA and 5G-AC-SA, with the difference that the constraint of having a sum of priorities equal to 1 is extended to all users in the system.
The parameters of our simulation model have been listed in Table I and Table II [15].

Fig 3 indicates the average throughput for UEs from different slices. It can be noticed that 4G-SA allocates the minimum data rate to services with low requested data rate (i.e., web browsing and car management) in order to provide higher data rate to the TV streaming services. The reason behind this behaviour is that, to maximize the QoE, 4G systems prefers to boost the performance of users with higher QoS requirements. It is worth noticing that, with 4G-SA, users belonging to TV streaming A and B experience the same throughput, although these such services belong to two different slices with different priorities. This strongly underlines that 4G-SA is not able to guarantee prioritization on a slice-basis, but only on a service-basis. Our proposed approaches, meaningfully increases the data rate for all the service slices compared to 4G-SA. We can observe that 5G-SA-AC guarantees higher throughput compared to 5G-SA, by thus highlighting the importance of the proposed admission control in achieving better utilization of spectrum resources. Finally, we can note that both 5G-SA and 5G-SA-AC introduces differentiation in the throughput of slices TV streaming A and B, as they are designed to take into consideration the inter-slice priority.

Fig 4 indicates the average QoE level of different slices. It can be noticed that 4G-SA provides the lowest QoE level for different slices compared with our proposed algorithm. In addition, we can note that although 4G-SA guarantees the same throughput to TV streaming A and TV streaming B, the QoE of streaming B is lower than the one of TV streaming A because of the fact that TV streaming B has a lower priority level compared to TV streaming A. When focusing on 5G-SA and 5G-AC-SA, we can observe the following: (i) they substantial increase the QoE compared to 4G-SA; (ii) the offer a better fairness in the QoE experienced by the users of different slices. Indeed, when focusing on the QoE values of TV streaming A and Web browsing, we can observe that 5G-SA and 5G-AC-SA guarantee a lower the difference in the QoE for these slices compared to 4G-SA. This means that our approaches are also able to guarantee a better fairness compared to 4G-SA.

Fig 5 indicates the time averaged total QoE level by considering a varying number of UEs per slice. From this figure, we can observe that the overall QoE for 4G-SA decreases as the number of UEs increases. This is due to the fact that when the number of UEs increases, this algorithm tries to increase the overall data rate of the system by allocating resources to the services with higher data rates (as shown in Fig. 5). From a global point of view, this involves a QoE reduction as 4G-SA does not consider inter-service priority. Our approaches, on the contrary, are based on the idea of exploiting the inter-slice priority for slice allocation. As a consequence, our approaches do not show a meaningful degradation of the overall QoE when increasing the number of UEs. It is worth noticing the

### Table I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cells one macrocell</td>
<td></td>
</tr>
<tr>
<td>LTE bandwidth</td>
<td>5 MHz</td>
</tr>
<tr>
<td>UEs distribution</td>
<td>uniform</td>
</tr>
<tr>
<td>Overall number of UEs</td>
<td>200</td>
</tr>
<tr>
<td>Overall interval</td>
<td>10s</td>
</tr>
<tr>
<td>Scheduling frame</td>
<td>10ms</td>
</tr>
<tr>
<td>TTI Duration</td>
<td>1ms</td>
</tr>
<tr>
<td>Noise Spectral Density</td>
<td>-174 dBm/Hz</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Slice Parameters</th>
<th>Th (kbps)</th>
<th>slice priority</th>
<th>4G priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV Streaming A</td>
<td>1000-1500</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>TV Streaming B</td>
<td>1000-1500</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>Car Management</td>
<td>400-700</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Web Browsing</td>
<td>100-300</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>
benefits introduced by 5G-AC-SA, that is able to guarantee a lower QoE decrease compared to 5G-SA.

Fig. 6 indicates the amount of free network resources after the resource allocation step.

Fig. 6. Percentage of free resources after the resource allocation step.

In this paper, we presented a novel approach for resource allocation in the 5G networks with network slicing. Our approach is a heuristic based prioritized admission control mechanism that takes into consideration both the inter- and the intra-slice priority and performs the resource allocation. Accordingly in order to meet the QoS requirements dictated by the service slice. Our approach increases the QoE experienced by mobile UEs as well as allows a better management of network resources.

ACKNOWLEDGMENT

This work has been supported in part by the 5GPP Virtual (Virtual and programmable industrial network prototype deployed in operational wind park) Project.

REFERENCES


VI. Conclusion

In this paper, we presented a novel approach for resource allocation in the 5G networks with network slicing. Our approach is a heuristic based prioritized admission control mechanism that takes into consideration both the inter- and the intra-slice priority and performs the resource allocation. Accordingly in order to meet the QoS requirements dictated by the service slice. Our approach increases the QoE experienced by mobile UEs as well as allows a better management of network resources.