Statement of Responses to the Editor and the Reviewers of Paper-TNSM

We would like to thank the editor and reviewers for their comments on our manuscript. We hope that the modifications that we have made to the manuscript, and the responses that we have provided herein will alleviate the reviewers' concerns. Below, please find our detailed responses to the editor and reviewers' comments and suggestions.

Editor
Comments to the Author "The paper has gone through three independent reviews and also checked by myself. The topic is interesting and there are merits. But there are some major concerns on the contribution, the assumption, the model derivation, and the experimental results. We hope the review comments are useful for further improving the quality of the paper. I'd therefore recommend the Resubmission"

Response:

We would like to thank the editor for his comment on our manuscript and giving us the opportunity to resubmit it. We have utilized the comments to improve our paper and eliminate the problems.

Reviewer 1
Comments to the Author " The paper proposes resolving the problem of resource allocation to network slices by using a new algorithm applied to the reformulation of the problem. The results seem promising but the paper has some issues. "

Response:

We would like to thank the reviewer for the careful and thorough reading of this manuscript. We hope that the responses provided herein can alleviate the reviewer's concerns.

Comment1: "First, the structure of the paper makes it somewhat hard to follow and there are some mistakes in the text. A proofread is required before it can be accepted for publication. "

Response:

We have changed the paper's introduction and checked the text entirely according to this comment.

Comment 2: "Although the paper demonstrates its claims, the relaxation of conditions from the original problem formulation is not well justified. It is not clear why the initial formulation of the problem is not feasible and the reformulated makes it feasible while retaining some level of quality of the solutions given. A deeper analysis of both formulations must be presented. "

Response:

The problem in (13) is feasible and has feasibility points that are discussed in subsection IV-B, and we introduce a fast algorithm (i.e., Algorithm 3) for feasibility points. Although the problem is feasible, it is not convex and difficult to solve. Since problem in (13) is mixed-integer nonlinear programming with two integer variables, the PRB assignment, e, and the number of VNFs in slice s, M_s, and the problem is NP-hard. Solving the problem is not trivial. To solve the problem by inspiring Stackelberg, we reformulate the equation in (13g) to reduce one of the variables (i.e., M_s) that can be solved after obtaining the rate of UEs. WE notice that M_s is similar to the followers in Stackelberg Competition and power and PRB assignment is similar to leader. So, the new problem has two variables of power and PRB assignment. This new problem is convex by relaxing the binary variable, the PRB assignment, and estimating the lower bounds in (15) because the objective function and constraints of the problem are convex and can be solved by the Lagrangian function. After obtaining the power of UEs and PRB assignment, we can obtain the achievable rate of each UE so we can find the optimal number of VNFs. We made changes in Subsection III-A, accordingly.

Comment3: "Moreover, the baseline and FBDR methods used in the comparison are not well introduced. They are vaguely linked to related work but not as needed. The paper must clarify the relation of the related work and the compared alternatives. The paper must also contextualize the proposal among the related work by comparing their qualities and/or performance. "

Response:

Two different methods are used to compare with the performance of the proposed method (IABV) and show the optimality of our approach. The first one is a baseline scheme, which uses random PRB allocation. So, the allocation of PRB to each UE is random when we have low interference, but in figures with high interference, we randomly assign just one RB to each UE. Also, the association of O-RU is carried out based on distance. It means that each UE is assigned to the nearest O-RU. The optimal power is obtained using the CVX of Matlab, which uses the successive convex approximation (SCA) method since the problem is convex. After achieving power and other parameters, the achievable rate will be obtained and the optimal number of VNF is achieved from Lemma (1). The second one is similar to the fixed BBU capacity and dynamic resource allocation (FBDR) algorithm proposed in [18]. In this work, we have services with different QoS that contain UEs, which is similar to tenants with different QoS that is introduced in [18]. So, we used an algorithm similar to FBDR adapted to our conditions for comparison. Instead of BBU in C-RAN, we have O-DU and O-CU in O-RAN. To use the FBDR method, we should consider the fixed BBU capacity. We assume that O-DU and O-CU have fixed sufficient capacity in our system model. Also, our mid-haul link (F1 link) has adequate capacity, so there will be no issue using the FBDR method by separating BBU to O-DU and O-CU with this assumption. In this method, PRB and power are dynamically allocated. The number of VNFs is obtained from the simulation. The UEs are associated with the O-RU based on the quality of their channels and the channel distance instead of using the greedy algorithm 1 (GAA algorithm) for O-RU assignment. The figures in [18] show that dynamic BBU capacity and dynamic resource allocation (DBDR) perform better than FBDR for the same priority area. We will also see that our proposed algorithm performs better than FBDR in the numerical result section.

We add this response in subsection IV-A.

Comment4: "Finally, the source of the values used for the parameters in the evaluation must be clarified. "

Response:

We refer to the following list of references in our numerical result for this comment:

[1] 3GPP-TS-36.104-V13.3.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception (Release 13)," 2016-03.
[2] 3GPP-TR-36.931-V13.0.0, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) requirements for LTE Pico Node B (Release 13)," 2016-01.
[3] 3GPP-TS-25.101-V4.13.0, "User equipment (UE) radio transmission and reception (FDD)(release 4)," 2006-12.
[4]E. Mohyeldin, "Minimum technical performance requirements for imt-2020 radio interface(s)," 2020.
[5] ETSI-TR-138-913-V14.3.0, "5G; study on scenarios and require- ments for next generation access technologies(3GPP TR 38.913 version 14.3.0 release 14)," 2017-10.

In [1 page 11], the amount of BW is explained. In [2], the noise power is represented. The amount of delay for eMBB and URLLC are listed on Pages 24 and 25 in [5]. Also, on Page 26, the mMTC packet size is stated. On Page 25, the URLLC packet size is stated. In [3], on Page 13, Table 6.1, the maximum power is stated. In [4], on Page 4, the delay of URLLC and eMBB is and also the spectral efficiency of eMBB is noted.

We add these references in IV-A (references [41]-[45]).

Reviewer 2

Comments to the Author: "The paper focus on the aspect of network slicing in 5G cellular network which entails a service aware resource allocation of the different required virtual network functions (VNFs) for different slices which have different characteristics. More specifically, the paper proposes a mixed integer mathematical problem which in the original form is non-linear and hence hard to solve. To tackle this challenge the optimization problem is decomposed into two sub-problems where the solutions are not optimal however numerical investigations show that the solutions are competitive. In general, the paper is well written and structured."

Response:

We would like to thank the reviewer for careful and thorough reading of this manuscript and for the thoughtful comments and constructive suggestions, which help us to improve the quality of this manuscript. We hope that the responses provided herein can alleviate the reviewer concerns.

Comment 1: "In terms of taking the actual delay in the proposed there are some concerns which might be important for some time critical applications especially in the ultra-reliable low latency communications (URLLC) but also for different applications that fall under the enhanced mobile broadband (eMBB) generic service framework. First of all, in the paper the authors only consider processing delay and ignore propagation and transmission delay. Since multiple paths in reality could be utilized the role of the above two components might play an important role. Nevertheless, the authors need to clearly mention why those two components are not considered, saying for example are constant is not a good enough reason to be ignored since as eluded those change based on routing decisions."

Response:

Taking propagation and transmission delays into formulations is straightforward, but we avoided that for the sake of better presentation. However, they can be added to the system model easily. In this paper, we focus only on the processing delay to find the optimal number of VNFs, and we consider the other two delays are fixed; In the following, we describe more about these two types of delays.

The propagation delay is considered when each O-RU can connect to the number of O-DUs and can select to connect to which O-DU based on the route, capacity, and priority. Also, the connections of O-DUs to O-CUs is the same as O-RU to O-DU. Therefore, different routes are added to the system to solve the routing problem. However, it provides a new system model. So, the connection between O-RUs to O-DUs is fixed and transparent. As a result, the propagation delay is fixed and does not affect the optimization problem. The following is a brief calculation of propagation delay. Since the distance between the O-RU to O-DU is about 10 km and also the distance between O-DU and O-CU is about 80 km. Moreover, the distance from O-CU to the

network should not exceed 200 km [**?**]. so, the propagation delay is about $T^{\text{pro}} = (10+80+200) \times 10^3/(3 \times 10^8) < 1ms$. Since fronhaul, midhaul, and backhaul are fiber optics, c is the speed of light. Also, due to the edge technique in O-DU or O-CU for users with low latency, this amount of latency is greatly reduced. But we also do not focus on edge processing in this paper. You can find more about these distances in this link. The following is a brief calculation of transmission delay to show that it does not affect the optimization since it has a small amount. In URLLC and mMTC, the mean packet size can be between 20 to 32 byte; Also, the minimum data rate is assume to be $46bits/sec/Hz \times BW(180KHz)$. So the transmission delay from O-RU to O-DU is about $T^{fr,t} = \frac{20\times8}{46\times180\times10^3} < 2us$. As a result, the $T^{fr,t} \approx T^{mid,t} \approx T^{b,t}$. for eMBB, the packet size can be 100 times larger and the delay is not exceed the 0.6ms.

We add this response in subsection II-D.

Comment 2: "Also, note that for calculating processing delay requests for different services arriving at blade servers for vnf applications might get different treatment on how they access VMs or containers hence a single queue with non-priority and/or preemption (m/m/1) might not be a good approximation on the performance."

Response:

The processing delay can be modeled as an M/M/1 queue; since the number of arrival packets in the system are from a large number of independent sources. Moreover, the impact of a single packet on the system's performance is minimal. Also, the queue discipline will be first-in first-out (FIFO), and the arrival packet is assumed to follow a Poisson process. The system's clock is constant, and the size of the tasks is not fixed. So, we suppose that service times are exponentially distributed [36]. In addition, we assume that the arrival packets of each service have the same priority, and we consider priority between services, not between UEs of one service. Because we assume that the UEs of a service have the same priority, their sent packets also have an equal priority. Also, the services are isolated, and this priority does not invalidate the queue assumption. Furthermore, one service's priority over another can be higher, affecting the whole optimization and not queue formulation since the UEs in each service have the same priority, and each service has its processing delay independent of other services.

We add this response in subsection II-D-1.

Comment 3: "Also worth noting, that later on in the problem formulation there is the notion of service priority δ_s for data rate but this seems not to be used for accessing cloud resources. For example one could have changed the optimization problem and considered average allocated transmission rate and optimize with the same priorities access to cloud resources (instead of allocating equal access to vnf resources)."

Response:

In this problem, since the priority is in the objective function of the equation in (13), the priority

has whole effect optimization. So, priority is given to the whole optimization problem, and the optimization problem obtains the number of VNFs by affecting the priority term. If we were to discuss the placement of VNFs on the data centers, the problem would become a knapsack or bin-packing problem, and here the prioritization would affect the resource allocation algorithm, including memory, RAM, CPU, and bandwidth. So the algorithm was implemented based on prioritization.

Comment 4: "Some form of rationalization should be given for abstracting Interference as Gaussian noise. This is important because we expect the system to be interference limited and cell edge users to experience significant different levels of performance compared to centre cell users."

Response:

$I^k_{r,u(s,i)}$ is the sum of the power of interfering signals and quantization noise represented as follow

$$I^k_{r,u(s,i)} = \tag{1a}$$

$$\underbrace{\sum_{\substack{l=1\\ l\neq i}}^{U_s} e^k_{u(s,i)} e^k_{u(s,l)} p^k_{u(s,l)} \sum_{\substack{r'=1\\ r'\neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(s,l)} g^{r'}_{u(s,l)}|^2}_{\text{(intra-slice interference)}} + \tag{1b}$$

$$\underbrace{\sum_{\substack{n=1\\ n\neq s}}^{S} \sum_{l=1}^{U_s} e^k_{u(s,i)} e^k_{u(n,l)} p^k_{u(n,l)} \sum_{\substack{r'=1\\ r'\neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(n,l)} g^{r'}_{u(n,l)}|^2}_{\text{(inter-slice interference)}} \tag{1c}$$

$$+ \underbrace{\sum_{j=1}^{R} {\sigma_q}^2 |\boldsymbol{h}^k_{r,u(s,i)}|^2}_{\text{(quantization noise)}}, \tag{1d}$$

where $e^k_{u(s,i)}$ is the binary variable to show whether the k^{th} PRB is allocated to the UE i in slice s, assigned to r^{th} O-RU. Furthermore, there is no inter-slice interference since slices are isolated and there is just intra-slice interference.

Here we have two Gaussian noise types: additive Gaussian noise and the other is Gaussian quantization noise. The second noise is added to the interfering signal and shown with $I^k_{r,u(s,i)}$ and it is different with interference. To obtain SNR as formulated in equation (**??**), let $y_{u(s,i)}$ be the received signal of UE i in s^{th} service formulated as

$$y_{u(s,i)} = \sum_{r=1}^{R} \sum_{k=1}^{K_s} \boldsymbol{h}^{H\,k}_{r,u(s,i)} g^r_{u(s,i)} e^k_{r,u(s,i)} x_{Q^k_{r,u(s,i)}} + z_{u(s,i)}, \tag{2}$$

where $x_{Q^k_{r,u(s,i)}} = x_{P^k_{r,u(s,i)}} + \boldsymbol{q}_r$. Also, $x_{P^k_{r,u(s,i)}} = \boldsymbol{w}^k_{r,u(s,i)} p^{k\,\frac{1}{2}}_{r,u(s,i)} x_{u(s,i)}$, and $x_{u(s,i)}$ depicts the transmitted symbol vector of UE i in s^{th} set of service, $z_{u(s,i)}$ is the additive Gaussian noise

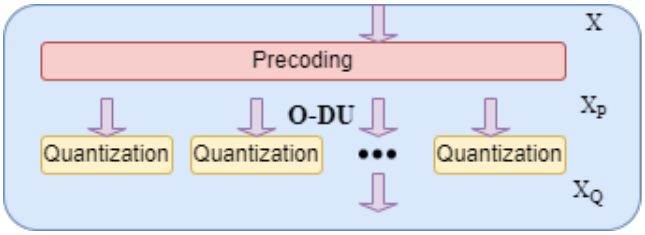

Figure 1: Precoding and Quantization of Signal

$z_{u(s,i)} \backsim \mathcal{N}(0, N_0)$ and N_0 is the noise power. Moreover, x_P denotes the precoded message before compression, and x_Q illustrates the precoded message after compression that is shown in figure 1. In addition, $\boldsymbol{q}_r \in \mathbb{C}^J$ indicates the quantization Gaussian noise ($\boldsymbol{q}_r \backsim \mathcal{N}(0, {\sigma_q}^2 \boldsymbol{I_R})$), which is made from signal compression in O-DU.

We add this response in subsection II-B.

Comment 5: "Constraint 13n (changed to 13k in new version) is not clear - it means that there are VMs and/or containers available in the network and an operator denies service due to an energy consumption budget (which again is not very detailed to capture the actual energy consumption of each node/vnfs under different loads etc.). In general the subsection on VNF power consumption is very limited in scope. "

Response:

A significant issue facing the industry is reducing energy consumption. Data centers are one of the most energy-consuming parts of a network. As a result, restrictions are placed on data centers' energy, including virtual machines (VMs). So, one of our goals is to limit the energy consumption of total VNFs that can be run as VM on data centers. So, by applying a custom policy on total power consumption, we can control data centers' power consumption ($\phi^{\text{tot}} \leq \phi^{max}$).

We add this response in subsection II-E.

Comment 6: "Some rationalization is needed on why the packet size is considered to be 20 bytes. "

Response:

We refer to the following reference in our numerical result for this comment:

[1] ETSI-TR-138-913-V14.3.0, "5G; study on scenarios and require- ments for next generation access technologies(3GPP TR 38.913 version 14.3.0 release 14)," 2017-10.

On [1, Page 25], the URLLC packet size is stated to be 32 bytes, and on Page 26, the packet size of mMTC is stated to be 20 bytes. We changed the URLLC packet size to 32, but it has little impact on the simulation result.

We add this reference in IV-A (reference [45]).

Comment 7: "Comparison with [18] might not be fair since that work also considers BBU capacity and also performs admission control functionalities, also there are different tenants that have different users with variable required QoS. "

Response:

We use a method which is similar to the fixed BBU capacity and dynamic resource allocation (FBDR) algorithm proposed in [18]. In this work, we have services with different QoS that contain UEs, which is similar to tenants with different QoS that is introduced in [18]. So, we used an algorithm similar to FBDR adapted to our conditions for comparison. Instead of BBU in C-RAN, we have O-DU and O-CU in O-RAN. To use the FBDR method, we should consider the fixed BBU capacity. We assume that O-DU and O-CU have fixed sufficient capacity in our system model. Also, our mid-haul link (F1 link) has adequate capacity, so there will be no issue using the FBDR method by separating BBU to O-DU and O-CU with this assumption. In this method, PRB and power are dynamically allocated. The number of VNFs is obtained from the simulation. The UEs are associated with the O-RU based on the quality of their channels and the channel distance instead of using the greedy algorithm 1 (GAA algorithm) for O-RU assignment.

We add this response in subsection IV-A.

Comment 8: "Also, interference is measured in a more detailed manner (maximum interference per UE) and hence when this relaxed (Guassian noise) the performance expected to slightly increase. Hence, some more detailed discussion on what has been assumed is needed. "

Response:

Here we have two Gaussian noise types: additive Gaussian noise and the other is Gaussian quantization noise which is shown in Fig 1. We offer the sum of interference and the Gaussian quantization noise with $I^k_{r,u(s,i)}$. Also, the Gaussian quantization noise is independent of interference and related to channel gain of UEs.

We add this response in subsection II-B.

Reviewer 3
Comments to the Author "The authors propose a resource allocation scheme for network slicing in an Open RAN scenario. They consider three network slice types namely eMBB, URLLC and mMTC and provide a solution for end-to-end slicing considering resource allocation over the RAN domain following the proposed ORAN architecture, as well as VNF allocation. In general, the paper is well written, however some parts need to be rephrased and restructured."

Response:

We would like to thank the reviewer for careful and thorough reading of our manuscript and for the thoughtful comments and constructive suggestions, which helped us to improve the quality of this manuscript. We hope that the modifications that we have made to the manuscript, and the responses that we have provided herein will alleviate the reviewer concerns.

Comment 1: "The paper presents an interesting solution and an extremely well formulated mathematical problem; however, the main contribution of the paper is hard to grasp. For instance, the introduction of the paper is very generic. There are a lot of concepts and methods explained, nonetheless not related to the proposed solution. The main issue lies in the structure of the work. The main contribution only appears at the end of page 2, where after an extensive reading the interest of the reader starts to vanish. I would definitely suggest a restructuring here. For instance, directly hint the main objectives and motivation of the work to prepare the reader for what is following. Moreover, the Related work could be a section of its own. In that way, the organization of the paper is clearer and easier to read. "

Response:

Thank you for the comment. We modified the introduction of the paper.

Comment 2: "Finally, a better distinction of the current proposal from the state-of-the-art is mandatory, otherwise it is hard to understand how the proposed algorithm differs from existing works in the literature which seem to provide solution to a similar problem."

Response:

The purpose of this paper is twofold. First and foremost, it is to design a system in the O-RAN structure with three types of services, namely, eMBB, URLLC, and mMTC. Simultaneously, it maximizes the total achievable data rate and meets the conditions of URLLC service low latency in the presence of numerous IoT devices requiring low power, leading to RAN slicing. Second, to model the delay for URLLC systems, we deal with the problem of obtaining the optimal number of VNFs in different layers of the O-RAN system. In this paper, we would like to optimize baseband resource allocation, i.e., power allocation, PRB allocation, O-RUs association, and VNF activation, to develop an isolated network slicing outline for different types of services in an O-RAN

platform. We use mathematical methods to decompose and convexify the problem and solve it using hierarchical algorithms to achieve these purposes. Unlike other papers, we concentrate more on the multiservice resource management of the RAN slicing in the openness and flexible O-RAN architecture. The novelty of this paper is to enhance the resource utilization of the overall wireless O-RAN system in the presence of the three generic service types introduced in 5G using RAN slicing. We also convexify and solve complex problems using mathematical concepts and obtain optimal resources.

Comment 3: "Furthermore, while the math introduced in the paper is solid, it is also hard to follow for a reader if illustrations are not presented. It would be easier if a Figure is introduced for instance to explain equations from 3a -3d, where a lot of variables are presented. Especially, for the concepts of inter and intra slice isolation, which are very crucial. Following, that logic more elaboration especially with respect to inter slice isolation and why it needs to be considered in the equation, is important, as one could say that a careful scheduling has to definitely avoid distribution of the same resources within a slice to different users i.e., (orthogonality constraint)."

Response:

We added this part in the system model. Assume there are K physical resource blocks (PRBs) in the system. Suppose each slice s consists of $\bar{K}_s$ preallocated virtual resource blocks that are mapped to Physical Resource Blocks (PRBs). Therefore, we have $\sum_s \bar{K}_s \leq K$.

Interference occurs when a UE in O-RU r using PRB k experiences an interference signal from other O-RUs in the set of $r' \in R \notin r$. UEs in each slice interfere in two ways in the technique of network slicing: the first is inter-slice interference between UEs of different slices, and the second is intra-slice interference between UEs of the same slice that is shown in figure 2. Network Slicing methods significantly reduce inter-service interference. There are some techniques to remove inter-slice interference. One of these techniques is to have two-time scale scheduling. The PRB scheduling to the slices is performed on the first time scale, and in the second time scale, the PRB scheduling to the UEs of slices is carried out. Since there are limited resources, inter-service interference cannot be eliminated entirely. The other method is to allocate part of the RB of eMBB services to URLLC and mMTC [2], [6], [34]. In this paper, we assume that the PRB scheduling is performed. Also, in subsection II-F1, we briefly study the PRB scheduling between slices. $I^k_{r,u(s,i)}$ is the sum of the power of interfering signals and quantization noise represented as follow

$$I^k_{r,u(s,i)} = \tag{3a}$$

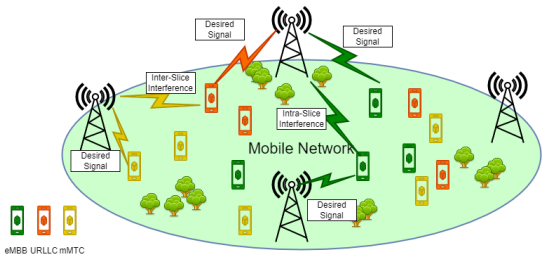

Figure 2: Type of Interference Signal

$$\underbrace{\sum_{\substack{l=1\\l\neq i}}^{U_s} e^k_{u(s,i)} e^k_{u(s,l)} p^k_{u(s,l)} \sum_{\substack{r'=1\\r'\neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(s,l)} g^{r'}_{u(s,l)}|^2 +}_{\text{(intra-slice interference)}} \tag{3b}$$

$$\underbrace{\sum_{\substack{n=1\\n\neq s}}^{S} \sum_{l=1}^{U_s} e^k_{u(s,i)} e^k_{u(n,l)} p^k_{u(n,l)} \sum_{\substack{r'=1\\r'\neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(n,l)} g^{r'}_{u(n,l)}|^2}_{\text{(inter-slice interference)}} \tag{3c}$$

$$+ \underbrace{\sum_{j=1}^{R} {\sigma_q}^2 |\boldsymbol{h}^k_{r,u(s,i)}|^2}_{\text{(quantization noise)}}, \tag{3d}$$

where $e^k_{u(s,i)}$ is the binary variable to show whether the k^{th} PRB is allocated to the UE i in slice s, assigned to r^{th} O-RU. Furthermore, there is no inter-slice interference since slices are isolated and there is just intra-slice interference.

We add this response in subsection II-B.

In this section, we have a brief study on the problem of PRB scheduling to eliminate the inter-slice interference and guarantee the isolation of slices [37]. We need to have an algorithm to remove the inter-slice interference before solving the problem in **??**. Firstly, we should assign PRBs to slices, and in the second step, the assignment of PRBs of each slice to each UEs of a specific slice is performed. So, the assignment of PRB can be completed in two steps to remove inter-slice interference and isolate the slices. Firstly, we assign PRBs to slices. Secondly, we allocate power of UEs, assign PRBs of slices to UEs, find the optimal number of VNFs for each slice and assign O-RU to UEs, which uses Algorithm **??**. Suppose $\mathcal{R}_s^{min}$ and $\mathcal{R}_s^{max}$ are the minimum data rate and maximum data rate of each UE in slice s, respectively. Firstly, we need to find the average PRB number used by UEs in each service. Since mMTC and URLLC transmit a short packet, each UE in mMTC and URLLC requires 1 PRB. So if slice s serves mMTC or URLLC services, with U_s UEs, it requires $K_s = U_s \times 1$ PRBs. For eMBB, assume the average rate of each UE in slice s serving eMBB UEs is $\bar{R}_s = Blog_2(1+\bar{\rho}_s)$, where, $\bar{\rho}_s$ is the average SNR of UEs in slice s. (eMBB slice) So, the minimum number of PRB that slice s with U_s UEs requires is $K_s^{min} = \lceil U_s \times \frac{\bar{R}_s}{\mathcal{R}_s^{max}} \rceil$. K_s^{min} is

the minimum number of PRBs needed for slice s, and K is the total number of PRBs in the system. Moreover, the maximum number of PRB that slice s with U_s UEs requires is $K_s^{max} = \lceil U_s \times \frac{\bar{R}_s}{\mathcal{R}_s^{min}} \rceil$. K_s^{max} is the maximum number of PRBs needed for slice s, and K is the total number of PRBs in the system. Also, $K_s = (K_s^{min} + K_s^{max})/2$ is the average number of required PRB in slice s (eMBB slice). Our goal is to obtain the number of PRBs assigned to each slice s ($\bar{K}_s$). The problem can be written as follow

$$\max_{\bar{\boldsymbol{K}}_{\boldsymbol{s}}} \quad \sum_{s=1}^{S} \delta_s K_s \ln(\bar{K}_s) \tag{4a}$$

$$\text{subject to} \quad \sum_{s} \bar{K}_s \leq K \tag{4b}$$

$$K_s^{min} \leq \bar{K}_s \leq K_s^{max} \quad \forall s \in S_1, \tag{4c}$$

$$\bar{K}_s \leq K_s \quad \forall s \in S_2, S_3, \tag{4d}$$

We use logarithms to assign PRBs to all slices to make them equally fair [37]. Equation (14b) illustrates that the sum of PRBs of slices can not exceed the maximum number of PRBs. Equation (14c), restrict the number of PRBs of eMBB slices and (14d), limit the number of PRBs of URLLC and mMTC slices By relaxing $\bar{K}_s$, the objective function and constraints become convex and can be solved using the Lagrangian function.

Also, we added a new subsection II-E-1 discyussing PRB scheduling for this question.

Comment 4: "Additionally, since the authors claim their novelty on the introduction of ORAN architecture, it becomes of utmost importance to consider concepts such as the creation of a slice, management of a slice and well as deletion, which bridge the mathematical framework to the practical one. For instance, how is a network slice created in the proposed work? How are the requirements of a slice fed to the algorithm? How is the monitoring of a slice performed?"

Response:

This subsection aims to examine slice management, including creating, managing, and deleting slices. Network slices generally have four life cycle stages: preparation, commissioning, operation, and decommissioning [38].

- Preparation phase: The network slice instance (NSI) does not exist in the preparation phase. In this phase, operators plan to create an NSI, such as designing the NSI template, onboarding users, and preparing the environment. Also, the evaluation of requirements is performed in this step.

- Commissioning phase: In the commissioning phase, the creation of the NSI is done. In this phase, the requirements are considered and allocated to the slice.

- Operation phase: During the Operation phase, NSIs are activated, managed, monitored (e.g., KPIs), modified, and deactivated. As the slice enters the activated phase, it is ready to support services, and as the slice exits the de-activated phase, the slice is inactive, and communication services are stopped.
- Decommissioning phase: A NSI that is decommissioned no longer exists after the decommissioning phase.

We added a new subsection II-G about Slice management for this question.

Comment 5: "Finally, the authors propose an interesting solution, however in none of the results information with respect to the convergence time of the algorithm were presented. This becomes extremely crucial when considering the real deployment of such a solution. In that regard, some findings with regard to this aspect need to be definitely included in the work."

Response:

It is shown in Fig. 11 and subsection III-C-1 and III-C-2.