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

We would like to thank the editor and reviewers for their constructive comments on our manuscript. It has been beneficial in revising this paper, and we have improved both the technical content and presentation quality through their assistance. We greatly appreciate their generous help. Moreover, we have reviewed and incorporated all the comments and suggestions. We hope that the modifications we have made to the manuscript and the responses 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 "I think the paper should undergo a major revision. It is important to take special attention to the comments of reviewer 2."

Response:

We would like to thank the Editor for his comments and concluding the revisions on our manuscript and for giving us this opportunity to improve our paper. We have used the comments to improve our paper and eliminate problems.

Reviewer 1
Comments to the Author " The authors have addressed the comments provided in a previous review. The contributions are clear and the results are compared to some related work, so the reader can somewhat contextualize the proposed solution. "

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: "However, the paper still needs some improvement in terms of readability. Some paragraphs are too long "

Response:

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

Comment2: " the figures are difficult to locate from their reference in the text (figures at top, as typical IEEE template is better), the text in the figures is too small "

Response:

Thank you for this comment. We changed the paper's figures totally according to this comment.

Comment3: "the contributions discussed in the introduction are not easily found in the other parts of the paper. "

Response:

We thank the reviewer for adding clarity to our paper and reducing ambiguity in this condition. We changed the contribution in a better way based on this comment. We have removed some sentences and rewritten the contribution in order to our paper's section. Below we write the whole contribution section.

The purpose of this paper is twofold. First and foremost, our goal 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 enhance the resource utilization of the overall wireless O-RAN system and 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. We also convexify and solve complex problems using mathematical concepts and obtain optimal resources.

In this paper, as depicted in Figure 1, the downlink of the O-RAN system is studied. The main contributions of this paper are summarized as follows:

- The paper presents a network slicing model for three 5G services: eMBB, mMTC, and URLLC. We examine the problem of radio resource allocation and VNF activation within the O-RAN architecture. Based on different types of services with different QoS and service priorities, we formulate a problem for allocating baseband resources to maximize the weighted throughput of O-RAN.

- The focus of our paper is on the multi-service resource management of the RAN, slicing in the flexibility, openness, and openness of the O-RAN architecture.

- We propose an algorithm for resource management in a two-step, with the first-step VNF activation, power allocation, PRB association, and the second-step O-RU association. In the first step, we reformulate and simplify the problem to find an upper and lower bound for the number of activated VNFs and use the Lagrangian function and KKT conditions to find optimal power and PRB allocation. For the second step, the problem of O-RU association can be converted to a multiple knapsack problem and solved by the Greedy algorithm.

- We talk about the initial point and the feasible region for the numerical results and introduce a fast algorithm that is less complex than our method to realize the feasible region for our problem.

- We perform numerical experiments to analyze the performance of the proposed algorithm, which proves to have a higher data rate than both the baseline scheme and data-driven method. Interestingly, our results show that this algorithm performs close to the optimal solution in low interference.

Reviewer 2
Comments to the Author " The Reviewer would like to thank the authors for their revision of the paper with respect to the comments and suggestions in the previous submission. Whereas the authors have considered parts of the suggestions and improved specific areas of the work, there are still a lot of remaining issues of concern in the current version of the paper. "

Response:

It is our pleasure to thank the reviewer for the careful reading of this manuscript, providing thoughtful comments, and offering constructive suggestions that strengthened and enhanced its quality. We hope that the responses provided herein can alleviate the reviewer's concerns.

Comment1: "While the authors have taken into consideration the Reviewer's comments about the re-structuring of the Introduction, I am afraid that the current version is not in the adequate state for publication. Initially, the paragraphs of the Introduction have no logical connection among them. It seems that until the Related Literature subsection all paragraphs are disconnected and only introduce concepts. "

Response:

We would like to thank the reviewer for the careful and thorough reading of our manuscript and the thoughtful comment and constructive suggestions that helped us improve the quality of this manuscript and make it more readable. We changed the introduction and add it below

Network slicing is the most effective solution for 5G wireless cellular networks to achieve the desired level of QoS (i.e., quality, delay, power, etc.). Network slicing can provide resource isolation for various services, increasing the system's efficiency. This architecture has several implementations, including core slicing, radio access network (RAN) slicing, and both for different 5G services. The network slicing concept has the potential to serve multiple services with varying architectures and quality of service (QoS) requirements.

Recent discussion of 5G wireless systems has been around three services, namely enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-to-machine communications (mMTC). Depending on QoS, each service requires its slice of the network. The eMBB service fulfills the demand for high capacity and throughput. Moreover, the URLLC service offers autonomous vehicles, tactile internet, remote surgery services, and other high-quality and low latency services. However, mMTC comprises a large number of internet of things (IoT) devices that transmit small payloads [1]–[6]. Nevertheless, the existing radio access network (RAN) architecture lacks adequate flexibility and openness to manage these demands for various services simultaneously. Therefore, development in RAN architecture is needed to support these requirements for different services simultaneously. Open radio access network (O-RAN) is

the new RAN generation introduced to deal with these issues.

The hardware is decoupled from the software in the O-RAN architecture, and each component is implemented as a virtual network function (VNF) that can be deployed on a virtual machine (VM) or container. Virtual network functions (VNF) are system function blocks in network function virtualization (NFV) systems. The concept of NFV refers to the separation of network software and hardware elements. Therefore network functions can run on commodity hardware. The NFV technology offers to execute VNFs as VMs or containers on a cloud environment [7], [8]. As a result, some O-RAN components defined in section III, such as user plane function (UPF), O-RAN central unit (O-CU), O-RAN distributed unit O-DU, and RAN Intelligent Controller (RIC)-near real-time, are near real-time virtualized and implemented as VNFs. VNFs can also be deployed as virtual machines (VMs) or containers.

This study presents a technique for creating isolated network slices outlines in O-RAN architecture to provide the specific QoS for eMBB, URLLC, and mMTC. As well as baseband resources, the number of VNFs is also taken into account to reduce latency, especially for URLLC services.

Comment2: " Moreover, I find the ORAN explanation in the Introduction very lengthy and unnecessary. I would definitely suggest to move that as a background Section. "

Response:
I would appreciate your comments to enhance and clarify this section of the paper. I moved the background as a section.

Comment3: " Furthermore, it is surprising how the Introduction does not contain a single research challenge. Why is the problem relevant at all? Why do you even need to study this problem and why is it hard to solve? "

Response:
We add the motivation subsection and answer these questions in it. Below we rewrite this section.

The studies in [9]–[11] have investigated resource allocation in C-RAN by considering the limitation on power and delay, respectively. However, this architecture is inefficient whenever we have different services with different QoS simultaneously. Additionally, RAN slicing requires a more flexible and open architecture. Therefore, we need a new architecture that supports slicing to implement RAN slicing for various services. O-RAN architecture has emerged as a new architecture that can serve different services simultaneously using RAN slicing. In [2], the RAN slicing is considered for C-RAN architecture for eMBB and URLLC. Nevertheless, the authors did not consider latency for URLLC, which is the main property of the URLLC. Moreover, the authors did not consider the weakness of C-RAN architecture. In [12], the authors examine the total delay of the UE in the O-RAN architecture. However, the paper did not consider the different services and other QoS.

Since 5G defines different services with different QoS, we need to analyze the resource allocation for each 5G service, using RAN slicing to guarantee the QoS for them. Consequently, we consider the O-RAN architecture, the new generation of RAN architecture that can implement RAN slicing for different 5G services. We investigate the problem of resource allocation in the O-RAN architecture for the three services defined in the 5G with different QoS serving simultaneously. Therefore, we want to study the problem of obtaining the optimal number of VNF, RB assignments, and power allocation to maximize the system's throughput and guarantee the QoS of services. This problem is mixed-integer non-linear programming. Hence, it is difficult to solve and requires some relaxation, convexification, and other methods for obtaining the sub-optimal solution presented in the following sections.

Comment4: " In turn, in the main contributions subsection, the initial sentences of almost all contribution paragraphs cannot be considered as contributions. For instance:
1) "We have carefully considered the processing delay and the VNF resources needed for the slice compared to other papers." 2) " Different services need to consider varying QoS conditions, including delay, power, and throughput." 3) "The main problem is mixed-integer non-linear programming that is extremely difficult to solve." "

Response:
We removed these sentences and rewrite the contribution. Below we write the changes.

The purpose of this paper is twofold. First and foremost, our goal 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 enhance the resource utilization of the overall wireless O-RAN system and 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. We also convexify and solve complex problems using mathematical concepts and obtain optimal resources.

In this paper, as depicted in Figure **??**, the downlink of the O-RAN system is studied. The main contributions of this paper are summarized as follows:

- The paper presents a network slicing model for three 5G services: eMBB, mMTC, and

URLLC. We examine the problem of radio resource allocation and VNF activation within the O-RAN architecture. Based on different types of services with different QoS and service priorities, we formulate a problem for allocating baseband resources to maximize the weighted throughput of O-RAN.

- The focus of our paper is on the multi-service resource management of the RAN, slicing in the flexibility, openness, and openness of the O-RAN architecture.

- We propose an algorithm for resource management in a two-step, with the first-step VNF activation, power allocation, PRB association, and the second-step O-RU association. In the first step, we reformulate and simplify the problem to find an upper and lower bound for the number of activated VNFs and use the Lagrangian function and KKT conditions to find optimal power and PRB allocation. For the second step, the problem of O-RU association can be converted to a multiple knapsack problem and solved by the Greedy algorithm.

- We talk about the initial point and the feasible region for the numerical results and introduce a fast algorithm that is less complex than our method to realize the feasible region for our problem.

- We perform numerical experiments to analyze the performance of the proposed algorithm, which proves to have a higher data rate than both the baseline scheme and data-driven method. Interestingly, our results show that this algorithm performs close to the optimal solution in low interference.

Comment5: " In a similar fashion, the authors have not considered the Reviewer's suggestion to move the Related Work as a separate Section in order to improve the flow and clarity of the Introduction. "

Response:
I changed the background from sub-section to the section part.

Comment6: " Regarding the Reviewer's comment with respect to the network slice management, whereas the authors have introduced the 4 stages of the life cycle of a network slice, the connection to their approach is very vague. It is still not clear how this process is achieved and how it is linked to their method. I was expecting at least a connection or introduction to Fig.1 where the system is shown. Additionally, more information with respect to the interaction with the system would have been required in that regard. "

Response:
We add the following part to the slice management section.

In the preparation phase, firstly, the evaluation of requirements is considered. Therefore, In

this phase, we need an algorithm to estimate the number of UEs and UE traffic in the system at different times. Moreover, based on this estimation, we need to evaluate resources, including the optimal number of VNFs for each slice, the optimal number of PRBs for each slice, and the analysis of power allocation. In this phase, we can use our algorithm after estimating the system's traffic. As shown in figure 1, we have three different slices for eMBB, URLLC, and mMTC. For each slice, the system must prepare MAC/RLC protocols for O-DU, PDCP/SDAP protocols for O-CU, UPF, SMF, and AMF. Moreover, O-RU, high PHY in O-DU, and O-CU-CP are shared between slices. Thus, we do not require estimating and preparing for the share environments and platforms in the network slicing cycles. After evaluating, assessing, and preparing the resources and environments for each slice, the commissioning phase is started. In this phase, the slices are created based on the previous phase. If the evaluation changes, the slice's resources can be modified. We can use our algorithm to assign resources to the services in the operation phase. If we need to remove a slice or any service is not used in a zone, and we need to remove the corresponding slice, the unshared resources are removed in the decommissioning phase.

Comment6: "With respect to the convergence time of the algorithm, the Reviewer had made a suggestion to the authors to include it in the current version of the manuscript, yet the authors do not seem to have done that. The authors' response points to Fig. 11. However if one looks at Fig.11 carefully, understands that only the Aggregate throughput is shown there. Even if one considers the Section III-C1 and III-C2 that the authors demonstrate as their convergence analysis and proof, there is no evidence or clarification of any delay value with respect to their algorithm's convergence. On the one hand it is very beneficial to have a mathematical proof that the algorithm converges. On the other hand, it is equally important to portray these findings with numerical values. In any case, this is not applicable in the current version of the manuscript. "

Response: