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

We would like to thank the editor and reviewers for their valuable comments on our manuscript. We have improved this paper's technical content and presentation quality through their assistance. We hope that the modifications undergone by the manuscript and the responses we have provided herein 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 that the paper has improved substantially. However, to make the paper suitable for publication, the comments of reviewer 2 should be taken into account."

Response:

Thank you for the constructive comments, which allowed us to improve our paper further. We are confident we addressed all the suggestions. Next, you can find our point-by-point response to the reviewers' comments and how the manuscript has been modified accordingly.

Reviewer 1
Comments to the Author " The authors have taken into consideration all the comments and suggestions raised by the Reviewer and have provided a substantial work to improve the quality of the manuscript. In that regard the Reviewer recommends the acceptance of the paper"

Response:

Thank you very much for your appreciation of our work and the valuable feedback we received from you.

Reviewer 2
Comments to the Author " The paper addresses most comments from previous revisions. However, its organization and presentation is still too weak. "

Response:
We thank the reviewer for taking the time to read the manuscript carefully, commenting thoroughly and offering suggestions that have made the manuscript stronger and more valuable. In this response, we hope to alleviate the reviewer's concerns.

Comment1: "First, the introduction is too long and does not synthesize the expectations of the rest of the paper. It must be changed to follow a the structure followed by other papers published in TNSM. "

Response:
Based on this comment, we have read several TNSM papers and reorganized the Introduction section accordingly. As a result, we shortened the Introduction section. Specifically, we revised it thoroughly and included an additional subsection about the organization of our paper. We trimmed extra parts and added more related literature reviews in the related literature. Additionally, we modified the background section by including some sentences/ideas that were originally in the Introduction section and reallocating others. Below we report the Introduction section.

1 Introduction

Network slicing is a key technology in 5G wireless systems. Specifically, it isolates network resources into slices, e.g., via core slicing and/or radio access network (RAN) slicing, for serving various services [1]–[3].

There are three main service classes in 5G, namely enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-to-machine communications (mMTC). Each service is assigned to a network slice depending on its corresponding quality of service (QoS) requirements. For instance, the eMBB service demands high capacity and throughput, e.g., 8K video streaming and immersive gaming. Meanwhile, the URLLC service provides ultra-reliable and low-latency connectivity, e.g., for autonomous vehicles, Tactile Internet, and remote surgeries. Finally, mMTC services require connectivity for a large number of Internet of Things (IoT) devices that transmit small payloads [4]–[6].

1.1 Motivation

The optimal resource allocation for the 5G systems is crucial to reducing costs and improving user equipment (UE). Significant challenges face these systems, including interference alignment, limited capacity of the fronthaul links, energy restrictions on virtual machines (VMs), etc [2], [7], [8].

Many studies have investigated resource allocation in cloud RAN (C-RAN) by considering a single service's power, data rate, and delay limitations. Unfortunately, the existing radio access networks (RANs) currently lack adequate flexibility and openness to handle these simultaneous service demands. Hence, a new RAN paradigm, called open RAN (O-RAN) architecture, has emerged. Therefore, O-RAN can simultaneously support multiple services at a lower cost by being flexible, layered, and modular. One of the fundamental problems lies in balancing services with different QoS, resource requirements, and priorities in O-RAN architecture [1], [9]–[11].

The purpose of this paper is to design a system in the O-RAN architecture to support the three types of 5G services, namely, eMBB, URLLC, and mMTC via network slicing and resource allocation.

1.2 Main Contributions

This paper studies the resource utilization of a downlink O-RAN system to develop an isolated network slicing outline for the three 5G services. We use mathematical methods to decompose and convexify the problem and solve it using hierarchical algorithms. The main contributions of this paper are summarized as follows:

- We examine the problem of baseband resource allocation, such as power, physical resource blocks (PRBs), O-RUs, and activating VNFs to maximize the weighted throughput of the O-RAN architecture. The three types of 5G service classes, i.e., eMBB, URLLC, and mMTC, are considered in this system. We take into account their corresponding QoS requirements and service priorities.

- We propose a two-step resource management algorithm for solving the optimization problem. In the first step, we reformulate and simplify the problem so as to find an upper and lower bound for the number of activated VNFs. Moreover, we use the Lagrangian function and Karush-Kuhn-Tucker (KKT) conditions to obtain the optimal power and PRB allocation. In the second step, the problem of O-RU association is converted to a multiple knapsack problem and solved by a greedy algorithm.

- We analyze the complexity of the proposed algorithms and demonstrate their convergence.

Additionally, we analyze the feasibility region of the problem and introduce a fast algorithm to check it numerically.

- We show via numerical results that the proposed algorithm outperforms two baseline schemes in terms of achievable data rate and mean total delay. Remarkably, the proposed algorithm performs close to the optimal solution in low-interference conditions.

1.3 Organization

This paper is organized as follows. Relevant literature related to our work is discussed in Section II, while Section III briefly overviews the O-RAN architecture. The system model and the problem formulation are described in Section IV and Section V, respectively. The details of our proposed resource management algorithm are introduced in Section VI. In Section VII, numerical results are provided to evaluate the performance of the proposed algorithm. Finally, Section VIII concludes the paper. For clarity, Table I lists the main acronyms used throughout the paper.

Comment2: "Second, many paragraphs are too long, what makes the path of the paper difficult to follow. For instance, paragraphs in the introduction give too much detail. Such details must be moved to the appropriate section and place a summary in the introduction. "

Response:

We have rewritten the introduction and literature review section, and reallocated some parts to more appropriate sections. We have shortened long paragraphs and sentences in this paper in order to make it easier to read. Furthermore, we modified the literature review and background too. We shortened the long paragraphs in the introduction, literature review, and background sections and added some additional literature reviews to make the paper more useful.

Comment3: "Regarding the technical content, the performance evaluation (Table III) must also include the related work, so that the reader knows the position of the proposed scheme in relation to dynamic resource allocation and baseline schemes. "

Response:

Based on this comment, we added the results for the baseline scheme and the DR methods in Table III in Section VII-C. Table III shows the execution time given a number of UEs for one service for the three methods. We run our simulation on the system with configuration (RAM = 8 GB, CPU = Core i5, SSD Hard Disk). As the number of UEs in the system increases, the execution time increases polynomially for all three algorithms. Since the baseline scheme is a simpler algorithm, with random PRB allocation and O-RU association based on distance, the execution time is less than the two other algorithms. Power and PRB are allocated in the DR scheme, but O-RUs are associated based on distance. Therefore the execution time is less than the proposed algorithm.

Table III: Execution Time vs. Number of UEs

Number of UEs	Execution Time (usec)		
	Proposed method	DR scheme	Baseline scheme
5	12.156	8.9546	6.6436
10	19.156	12.3112	8.7870
15	29.140	15.4778	9.5648
20	44.573	21.5342	14.8334
25	67.912	32.7926	21.5510

Comment4: "Moreover, it is not clear why the fact that a function increase justifies the convergence of the algorithm. "

Response:

We thank the reviewer for reporting this ambiguity. We revised this vague statement and added it to section VI-C-2. Below we reported the revised sentences.

Due to limited system resources, we have limits on VNFs' power, UE or O-RU power, fronthaul capacity, etc. As a result, the objective function, which is the aggregate throughput, cannot exceed its optimal value and become infinite. Therefore, if the aggregate throughput is infinite and increases without limit, the resources must also be unlimited. Hence, the system has an optimal solution: its maximum aggregate throughput in the feasible region.

Consequently, we can guarantee the convergence of the iterative algorithm if the objective function is the ascending function concerning the number of iterations and has an upper bound. Thus, it will converge to its optimum value if it is a strictly ascending function and to its local optimum if it is a non-monotonically ascending function.

Consider the aggregate throughput as $\mathcal{T}(\mathbf{P},\mathbf{E},\mathbf{G}) = \sum_{s=1}^{S}\sum_{i=1}^{U_s}\delta_s\bar{\mathcal{R}}_{u_{(s,i)}}$. In the first step of the iteration i of the algorithm **??** (IABV), we have $\mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^{i-1})$. In this step, optimal power and PRB allocation are obtained for the fixed O-RU association, so we have $\mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^{i-1}) \geq \mathcal{T}(\mathbf{P}^{i-1},\mathbf{E}^{i-1},\mathbf{G}^{i-1})$. In the second step of the iteration i, the optimal O-RU association is achieved to maximize the aggregate throughput. So we have this inequality $\mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^{i-1})$. As a result, we have $\mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^{i-1},\mathbf{E}^{i-1},\mathbf{G}^{i-1})$. Hence, in each step of the iteration, the aggregate throughput increased. Note that $\mathcal{T}^*(\mathbf{P}^*,\mathbf{E}^*,\mathbf{G}^*)$ is the achieved aggregate throughput for all the feasible resource allocation solutions of $\{\mathbf{P},\mathbf{E},\mathbf{G}\}$. So, $\mathcal{T}^*(\mathbf{P}^*,\mathbf{E}^*,\mathbf{G}^*) \geq \mathcal{T}(\mathbf{P}^i,\mathbf{E}^i,\mathbf{G}^i)$ and thus in each iteration, the aggregate throughput can not be larger than the optimal solution.

In addition, if we assume that the interference is set to be zero $I^k_{r,u_{(s,i)}} = 0$, and we suppose that each UE has the maximum power $p^k_{r,u_{(s,i)}} = P^{max}_s$, and we consider that all PRB is assigned to all

UE $e^{k}_{r,u_{(s,i)}} = 1 \forall s, \forall i$ and each UE is assigned to the nearest O-RU with the best channel quality. So, the solution of this allocation, is the upper bound for the aggregate throughput.

Thus, we can guarantee the local convergence of our iterative algorithm since the objective function $\mathcal{T}$ is the ascending function concerning the number of iterations and it has the upper bound which is not infinite.

Comment5: "In addition, only the convergence to the local optimum that is closest to the initial values is achieved, and it is not clear how it can be extended to global convergence. "

Response:
We agree with the reviewer that we only talked about the convergence of the proposed algorithm, and we do not talk about the global and local convergence.

Moreover, Fig. 12 shows the aggregate throughput vs. the number of UEs for the optimal solution and proposed algorithm. Therefore, there is a low interference comparison between the proposed method and the optimal solution for the system. In this figure, the proposed solution is near the optimal solution. Thus, the simulation demonstrates that it almost reaches the global optimum. Hence, Fig. 12 numerically illustrates the global convergence of the algorithm.

Meanwhile, Zangwill's global convergence theory of iterative algorithms is a standard theory for providing the global convergence of an iterative algorithm. We can use this theory in future extention of our work to prove our algorithm in a standard way, which is not straightforward in this paper.

Also, we added a short description of extending local convergence to the global convergence in the low interference system with low number of UEs. We added it in the Section VI-2 and below we reported it.

In addition, to extend our solution to the global optimum, we must prove that the algorithm monotonically increases in the non-optimal set of solutions and is Lipschitz monotone contraction mapping. Here, we briefly discussed our algorithm's global convergence for a low interference system.

In the low interference system, the PRB and VNF assignment is obtained straightforwardly, and the first step's problem is power allocation. In the first step of our algorithm with the fixed O-RU association, we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1}) > \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$ (strictly increase). Because in this step, the problem is reformulated to the power allocation, and the objective function is convex, and the convex functions are Lipschitz monotone contractions. However, in the second step of the algorithm, we can show as before that the algorithm is increased, but we can not talk about monotonically increases. Hence we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1})$. Nevertheless, the objective function is still Lipschitz monotone mapping. Accordingly, we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) > \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$. Therefore the objective function is the strictly ascending function concerning

the number of iterations and has an upper bound. Consequently, the algorithm converges to the global optimum solution in low interference.

Comment6: "In general, the paper must clarify to "what" is the algorithm converging or the global maximum can be understood by the reader and it can be confusing. "

Response:

We agree with the reviewer that we must clarify "what" the algorithm converges. In Section VI-C-2, we revised the convergence proved of the system and described that the algorithm converged to the local convergence. Moreover, as mentioned in the previous question, we briefly show that the algorithm can converge to a global convergence in the low interference system. Moreover, in Fig. 12, we numerically demonstrate our algorithm's global convergence in low interference and describe it in Section VII-C. Nevertheless, although we didn't demonstrate the unconditional global convergence of our algorithm, we did verify its unconditional local convergence.

Comment7: "Regarding presentation, there are some grammar typos, such as "reformulated as follow" in Section VI.B, second paragraph. "

Response:

We are pleased to thank the reviewer for improving the grammar in our paper. We have read the paper one more time and tried to fix any grammar problems in the paper.

Comment8: "Finally, according to IEEE style, tables and algorithms must be placed at top of the page, in addition to figures, which have already been correctly placed. "

Response:

It is our pleasure to thank the reviewer for this comment. We agree with the reviewer that the tables and algorithms need to be moved. As a result, all these mistakes were corrected.

Reviewer 3
Comments to the Author " The paper has addressed the reviewers' comments satisfactorily. "

Response:

We appreciate the reviewer's time and attention in reading this manuscript.

Comment1: "One minor issue is in Eq. (3) where the plus sign should be in the position after the quantization noise term. "

Response:

We would like to thank the reviewer for correcting our mistake. We modified the position of the plus sign, and the updated version is below.

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