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.

1. 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.

Introduction changed

1. In addition, 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.

Since problem \eqref{problem} is mixed-integer nonlinear programming with two integer constraints which are a binary variable ($e$ shows the PRB assignment) and an integer variable ($M\_s$ indicate the number of VNFs in slice s), the problem is NP-hard. Solving the problem is complex. To solve the problem by inspiring Stackelberg, we reformulate the equation \eqref{c16} to reduce one of the variables ($M\_s$) that can be solved after obtaining the rate of UEs ($M\_s$ is similar to 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 (PRB assignment) 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.

1. 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.

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 the lemma \eqref{lemma1}.

The second one is similar to the fixed BBU capacity and

dynamic resource allocation (FBDR) algorithm proposed in \cite{lee2018dynamic}. We assume here that the capacity of the O-DU (which is similar to BBU in the FBDR) is as large as the capacity of our fronthaul link so that we can use this method for our system model.

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 \ref{alg1} (GAA algorithm) for O-RU assignment.

The figures in \cite{lee2018dynamic} 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 next subsection.

1. Finally, the source of the values used for the parameters in the evaluation must be clarified.

<https://www.arib.or.jp/english/html/overview/doc/STD-T104v4_00/5_Appendix/Rel13/36/36104-d30.pdf> page 11 bandwidth

<https://www.arib.or.jp/english/html/overview/doc/STD-T104v4_20/5_Appendix/Rel13/36/36931-d00.pdf> page 10 noise power

<https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.03.00_60/tr_138913v140300p.pdf> page 24 embb delay urllc delay 0.5 ms

<https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.03.00_60/tr_138913v140300p.pdf> page 25 urllc latency 1ms

<https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Documents/S01-1_Requirements%20for%20IMT-2020_Rev.pdf> embb urllc latency, page 4 , peak spectral efficiency range of embb- page 7

<https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.03.00_60/tr_138913v140300p.pdf> page 27 packet size mmtc

<http://www.arib.or.jp/english/html/overview/doc/STD-T63v9_40/5_Appendix/Rel4/25/25101-4d0.pdf> table 6.1 power classes

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.

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.

We can reduce the complexity of our delay calculation and enhance the performance of our solution by eliminating propagation and transmission delays to focus on obtaining the optimal number of VNF.

In this paper, we focus only on the processing delays to find the optimal number of VNFs, and we consider the other two delays are fixed for simplification;

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 \cite{oranD1}. 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 use of the edge technique in O-DU or O-CU for users with low latency, this amount of latency is greatly reduced. But we do not focus on edge processing in this paper.

In URLLC and mMTC, the mean packet size can be between 20 to 32 byte; Also, the minimum data rate is assume to be $46 bits/sec/Hz \times BW (180 KHz)$. So the transmission delay from O-RU to O-DU is about $T^{fr,t} = \frac{20\times 8}{46 \times 180 \times 10^3} < 2 us$. 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.

[https://blogs.keysight.com/blogs/inds.entry.html/2020/06/30/5g\_testing\_what\_is-0nzj.html#:~:text=The%20data%20requirements%20are%20about,DUs%20is%20about%2080%20km.](https://blogs.keysight.com/blogs/inds.entry.html/2020/06/30/5g_testing_what_is-0nzj.html%23:~:text=The%20data%20requirements%20are%20about,DUs%20is%20about%2080%20km.)

1. 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.

The processing delay can be considered an M/M/1; since the number of arrival packets in the system is enormous, and the packets are independent. 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. We also suppose that service times are exponentially distributed. In addition, we assume that the arrival packets of each service have the same priority, and we consider priority between services. Because in this system, it is also assumed that the users of a service have the same priority, so their sent packages also have an equal priority over each other. Because the services are isolated, this priority within a service does not invalidate the queue assumption. Also, one service's priority over another can be higher, affecting the optimization.

1. Also worth noting, that later on in the problem formulation there is the notion of service priority \delta\_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).
2. 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.

Also, denotes the power of Gaussian additive noise, and I\_{r,u(s,i)}^{k} is the sum of the power of interfering signals and quantization noise. To obtain SNR as formulated in \eqref{eq1}, let $y\_{u(s,i)} $ be the received signal of UE $i$ in $s^{th}$ service formulated as

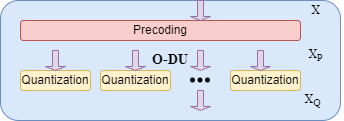
\begin{equation}\label{eq2}

y\_{u(s,i)} = \sum\_{r = 1}^{R}\sum\_{k=1}^{K\_s} \boldsymbol{h}^{H \: k}\_{r,u(s,i)} g\_{u(s,i)}^r e^k\_{r,u(s,i)}\mathfrak{y}^k\_{r,u(s,i)}+ z\_{u(s,i)},

\end{equation}

where $\mathfrak{y}^k\_{r,u(s,i)} =\boldsymbol{w}^k\_{r,u(s,i)}{p^{k \: \frac{1}{2}}\_{r,u(s,i)}} x\_{u(s,i)}+ \boldsymbol{q}\_{r}$, $ 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 $z\_{u(s,i)} \backsim \mathcal{N}(0,N\_0)$ and $N\_0$ is the noise power.

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.



1. Constraint 13n 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. (changed to 13k)

A significant issue facing the industry is reducing energy consumption. Data centers are one of the most energy-consuming. 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.

1. Some rationalization is needed on why the packet size is considered to be 20 bytes.

<https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.03.00_60/tr_138913v140300p.pdf> page 27 packet size mmtc

<https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol19_3/vol19_3_003en.pdf> packet size between 32 – 200 🡪 urllc

1. 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.

The second one is similar to the fixed BBU capacity and dynamic resource allocation (FBDR) algorithm proposed in \cite{lee2018dynamic}.

In this work, we have services with different QoS that contain UEs, which is similar to tenants with a different QoS. So, we used an algorithm similar to FBDR adapted to our conditions for comparison.

We assume here that the capacity of the O-DU (which is similar to BBU in the FBDR) is as large as the capacity of our fronthaul link so that we can use this method for our system model.

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 \ref{alg1} (GAA algorithm ) for O-RU assignment.

The figures in \cite{lee2018dynamic} 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 next subsection.

1. 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.

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\_{r,u(s,i)}^{k}$ and it is different with interference.

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.

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.

Introduction changed

1. 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.
2. 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).

UEs in different O-RUs can be assigned to different PRBs. As a result, a UE in O-RU $r$ using PRB $k$ may experience interference from other O-RUs in the set of $r' \in R -r$.

Network Slicing methods significantly reduce inter-service interference but do not eliminate it at all.

There are some techniques to minimize 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 \cite{alsenwi2021intelligent, setayesh2020joint, mei2021intelligent}.}

We have two types of interference, the first one is between slices (intra-slice) and the second one is between UEs in each slices or services (intra-slice) that is shown in figure \ref{fig:intf}

1. 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?
2. 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.

It is shown in figure 11 and In Iterative proposed algorithm Complexity order and convergence analysis is depicted