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.
2. 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.
3. 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.
4. 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.

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.

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.

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).

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.

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.

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

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.

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.

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.  
  
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.

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.

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).

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?  
  
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.