Base Station Popularity-Based Dynamic Resource Allocation for VNF

Sanshan Sun[1], Jianhong Zhou[1], Yao Sun[1], Gang Feng[1], Shuang Qin[1], Wenkui She[2]
[1] National Key Lab of Science and Technology on Communications,
University of Electronic Science and Technology of China;
[2] R & D Center, Sichuan Aostar Information Technology Co., Ltd.
Email: sanshansun@hotmail.com

***Abstract*—Network Function Virtualization (NFV) is emerging as a promising technology in future wireless communication system to achieve resource sharing through abstracting standardized network equipment into different types of virtual network functions (VNFs) to be placed in various network slices for diverse requirements. In order to satisfy the required QoS of serving users, each network slice deploys appropriate VNFs on different Base Stations (BSs) and orchestrates them for providing uniform service like an independent virtual network. However, due to the user mobility, the VNF on the BS may not have sufficient resource to provide QoS guaranteed services for newly accessed users. It is challenging to allocate adequate resource dynamically for VNFs to guarantee the required QoS of roaming users. In this paper, we propose a dynamic resource allocation scheme for VNF based on group mobility prediction. We first employ Markov chain and online learning method together to predict group mobility of users. Then we calculate the popularities of BSs for allocating more resource to the hotspot BSs (HBSs) with aim of enabling HBS permit the enormous service requests of approaching users. We propose a complementarity mechanism to maximize resource efficiency when implementing resource allocation for VNFs. Numerical results validate the effectiveness of our proposed dynamic resource allocation scheme.**

Keywords—Network Function Virtualization, Virtual Network Function, Group Mobility Prediction, Base Station Popularity, Dynamic Resource Allocation

I. INTRODUCTION

With the emergence and rapid development of new applications such as the Internet of Things and the Internet of Vehicles, existing mobile communication networks have been unable to meet the increasing communication needs of users in terms of system capacity, transmission rate and reliability. The fifth-generation mobile communication (5G) system has been considered as an promising paradigm to provide flexible service with its high frequency efficiency, energy efficiency, and elastic network architecture as well [1]. With the help of network function virtualization (NFV) technology, the dedicated network devices in traditional communication network are replaced by standardized servers in 5G system. The hardware resources (e.g., CPU and Memory) of the servers are abstracted into many types of basic virtual resources that can be further composed as a virtual network function (VNF) to provide specific service [2]. As employing software defined network (SDN) technology, 5G system can further orchestrate different VNFs to a service chain for serving users [3]. Specifically, the diverse service type results in different deployment scheme of VNF in the service chain. Moreover, the service chain is also called as network slice, which is been consider an independent virtual network for providing specific QoS guaranteed service [4] [5].

In general, the SDN-based controller executes resource allocation for VNFs at the time of implementing network slicing. Specifically, some network performance indicators such as QoS of service type, overall network load and available resource, would affect the allocated resource on VNFs, which further determines the service capability of network slice. After orchestrating various network slices based on existing service types, the sliced network would assign users to corresponding network slice according to the QoS of requested service. Moreover, the service area of network slice depends on the coverage of base stations (BSs) that have been deployed specific VNFs by virtualizing radio access network (RAN) device. Namely, if a network slice deploys some VNFs on a BS to provide service, the BS is covered by the network slice. Accordingly, the users who are served by specific network slice would be assigned a slice ID, and are capable of accessing network slice when approaching BSs that are covered by the network slice [6]. Hence, the network slice can provide ubiquitous services for mobile users.

In early network slicing scheme, in order to utilize resource efficiently, the virtual network resources are usually allocated in a global view of entire network. Specifically, taking into account of actual service load of each BS, little or less resource is allocated to the VNFs of BSs that have no service load. Apparently, this network slicing scheme mainly focuses on the static scenario, in which the resource allocation for VNF is respond to current serving users. However, due to the user mobility in mobile network, users accessing a certain network slice would fail to be served continually when moving into a BS that is not covered by the network slice. Hence, with the purpose of guaranteeing QoS of mobile users, the resource allocation for VNFs should fully consider the dynamic changes of service load on BSs.

Many researchers pay enormous attention to resource allocation for network slicing. Y. Zaki *et al*. [7] [8] present a framework for LTE virtualization through hosting multiple virtual eNodeBs on a single physical LTE BS and scheduling PRB among virtual eNodeBs. M. I. Kamel *et al*. [9] develop an efficient resource allocation scheme to allocate the radio resource blocks for LTE network slicing. The scheme keeps track of the service contracts with the SPs and the fairness requirements between cell-center users and cell-edge users. E. Pateromichelakis *et al*. [10] investigate the adaptive placement of Radio Resource Management (RRM) functionalities to the RAN nodes and the interactions among functionalities on per slice basis. Specifically, they consider that a central management entity should assign RRM to the controllers that manage the cluster of access nodes. K. Zhu *et al*. [11] study two-level hierarchical resource allocation problem, in which Infrastructure Provider (InP) abstracts the physical resources into isolated slices for each Mobile Virtual Network Operatore (MVNO) who then allocates the resources within the slice to its subscribed users. They eventually design a hierachical combinatorial auction mechanism to solve the two-level hierarchical resource allocation.

However, most literature consider less on the user mobility when implementing resource allocation for network slicing. Actually, user mobility would impact the resource utilization of network slice, particularly when the moving users lead to the change of the service load of some BSs that have been covered by specific network slice. Furthermore, the QoS of service running on a network slice would degrade when some BSs with high service load cannot provide sufficient resource to satisfy the demand of each user. Hence, we propose a dynamic resource allocation scheme based on the popularity of BS when implementing network slicing. On the basis of the prediction for user group mobility, we select those BSs that are likely to permit new accessed users as hotspot BSs (HBSs), and dynamically allocate sufficient resource for their VNFs.

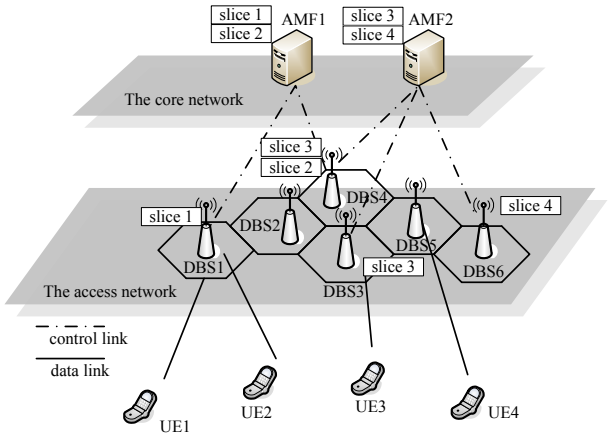
The rest of this paper is organized as follows. In Section II, we describe system model and user mobility model based on Markov chain, respectively. In Section III, we present the BS popularity based on group user mobility. In Section IV, we show the implementation for proposed dynamic resource allocation scheme. In Section V, we provide numerical results for performance evaluation of proposed scheme. In Section VI, we conclude the paper.

II. System Model and User Mobility Model

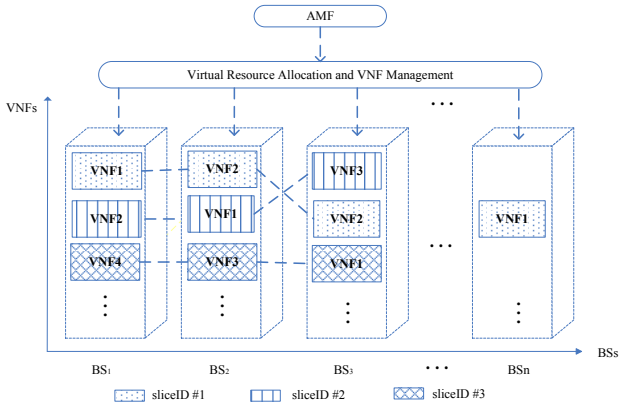
In this section, we introduce the system model that captures the characteristics of sliced network architecture. Meanwhile, we introduce the user mobility model based on Markov chain.

A. System Model

In system model, we consider a slice based mobile network as shown in Fig. 1 (a), which consists of the core network and the access network. The Access and Mobility Management Function (AMF) is running on the equipment of the core network to allocate virtual resources to different Virtual Network Functions (VNFs). Then AMF deploys these VNFs on some BSs of the access network to serve users. Due to the diversity of VNFs, the BSs are capable of providing diverse service for users to satisfy their personalized requirements. Moreover, the aggregation of different VNFs for providing specific Quality of Service (QoS) can be considered as a network slice. Hence, as shown in Fig. 1 (b), a network slice can cover multiple BSs, and each BS can be deployed multiple network slices simultaneously.



(a) Overview of the architecture for slice based mobile network



(b) Illustration of the VNF deployment on BSs

Fig. 1. System Model

We further consider that there are N BSs in the slice based mobile network, and the BSs deployed in a standard cellular fashion. We let $\mathrm{S}=\{BS_1,BS_2,...BS_n,...BS_N\}$ to denote the set of all BSs, and Ω_n to denote the set of neighbors for BS_n. Apparently, $\Omega_n \neq \varnothing$. Moreover, we consider that there are M UEs in the network, and $\mathrm{U}=\{U_1,U_2,\ldots U_m,\ldots U_M\}$ denotes the set of all UEs. When leaving from the serving BS, the UE should only move into one of the neighboring BSs. Specifically, as long as the neighboring BSs are deployed with the same slice as their assigned slice ID, the UEs will be served continuously during the move process.

B. User Mobility Model

Based on the above system model, each handover of the UEs can be deemed as an independent random transfer for UE from one BS to another. In particular, the current movement for UE is not affected by the previous track. Hence, we employ the Markov chain to model the user mobility.

Considering the basic elements of Markov chain are state space, initial probability and transition probability, we first construct the state space. As each UE is always be served by one specific BS before or after the handover, we regard one BS that serves UEs as one state. Hence, the states of Markov chain are N. Hence, denote by $\boldsymbol{p}_0=[\alpha_1\ \alpha_2\ \ldots\ \alpha_n\ \ldots\ \alpha_N]$ the initial probability for the UE to be served by each BS at the beginning of accessing the network. Specifically, α_n denote by the probability for UE to be served by BS_n. Apparently, $0\le\alpha_n\le 1$, and $\sum_{n=1}^{N}\alpha_n=1$. Meanwhile, we let $t_{i,j}$ to denote the direct transition probability for UE to move from BS_i to BS_j when executing handover. Therefore, the transition probability matrix for UE can be denoted by

$$\boldsymbol{T}=\begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,N} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ t_{N,1} & t_{N,2} & \cdots & t_{N,N} \end{bmatrix}, \tag{1}$$

where the element $t_{i,j}$ must satisfy the following constraints

$$0 \le t_{i,j} \le 1, \quad \forall i, j \in \{1,2,...,N\}, \tag{2}$$

$$\sum_{j=1}^{N} t_{i,j} = 1, \quad \forall i \in \{1,2,...,N\}, \tag{3}$$

$$t_{i,i} = 0, \quad \forall i \in \{1,2,...,N\}, \tag{4}$$

$$t_{i,j} = t_{j,i} = 0, \quad \forall j \notin \square_n. \tag{5}$$

Constraint (2) indicates that the transition probability for UE must be an integer between 0 and 1. Constraint (3) indicates that the UE must be provided service by specific BS as long as the handover is successful. Constraint (4) indicates that the UE cannot execute handover between the same BS. Constraint (5) indicates that the UE can only hand over to the neighboring BSs of his original serving BS.

We further let k to denote the times for UE to execute handover in the network, and $\boldsymbol{T}_k$ to denote the k-step transition matrix of Markov chain. Therefore, the k-step direct transition probability $\boldsymbol{p}_k = [\alpha_1^k \ \alpha_2^k \ ... \ \alpha_n^k \ ... \ \alpha_N^k]$ can be calculated by

$$\boldsymbol{p}_k = \boldsymbol{p}_0 * \boldsymbol{T}_k. \tag{6}$$

III. The Base Station Popularity Based on Group User Mobility

In this section, we first introduce the main idea of user mobility prediction based on online learning process. Then we introduce the popularity of BSs based on the group user mobility.

A. User mobility prediction

A. Mohamed *et al*. [12] propose a novel scheme for user mobility prediction based on online learning process. Inspired by their work, we utilize the real data of network operation from [13] to implement the prediction. Specifically, we sort and number the base station data and user data of a certain area, and integrate these data into the user movement path, and apply it to the online learning process shown in Fig. 2.

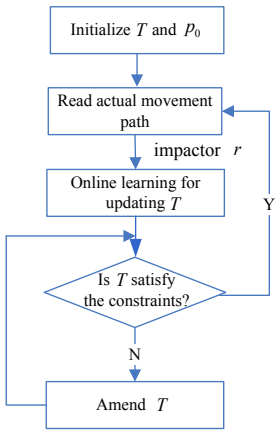


Fig. 2. Block diagram of online learning for user mobility prediction

As shown in Fig. 2, the online learning scheme for user mobility prediction consists of four steps. The details of each step can be summarized as the following:

1) Step 1: Initialize the parameters for the learning. We assume that the UE can access any BS fairly before executing handover. Hence, $\alpha_n = 1/N$. Moreover, we also assume that the UE should hand over to each BS fairly before the system start to learn from actual movement path. Therefore, the transition matrix $\boldsymbol{T}$ should be square with $N \times N$. We let $|\square_n|$ to denote the number of neighbors of BS_n. Then the element of transition matrix $t_{i,j} = 1/|\square_n|$, $\forall i \ne j$.

2) Step 2: Read actual movement path of UE for online learning process. Suppose a UE follows the actual movement path: $BS_a \to BS_b \to BS_c \to BS_d$. The path shows that the UE starts at BS_a, and arrives at BS_d through three handovers.

3) Step3: The online learning process starts to learn from the movement path of UE, and updates the transition matrix $\boldsymbol{T}$. Specifically, the number of times a handover occurs in the path determines the number of times the transition matrix is updated. For the supposed path in Step 2, the online learning process would update $\boldsymbol{T}$ in sequence for every handover. For instance, for first handover in the path $BS_a \to BS_b$, the process updates $\boldsymbol{T}$ according to the following:

$$t_{a,b}^1 = t_{a,b} + \sum_j t_{a,j} * r, \forall j \in \square_a, j \ne b, \tag{7}$$

$$t_{a,j}^1 = t_{a,j} - \frac{\sum_j t_{a,j} * r}{|\square_a| - 1}, \forall j \in \square_a, j \ne b, \tag{8}$$

where the rule (7) indicates that the direct transition probability from BS_a to BS_b increases if a real handover occurs between these two BSs. The increment of probability is provided by other transition probabilities. Meanwhile, the rule (8) indicates that the other BSs that execute no handover should reduce their transition probability. The reduction in each probability is equal to the averaging of the increment of increasing transition probability they contribute to. In (7), (8), the superscript "1" of $t_{a,b}^1$ indicates the number of times that handover occurs. The impactor r reflects the importance of current path for online learning. Specifically, r=1 indicates that the current path is decisive for updating transition matrix, and r=0 indicates that the update of transition matrix is not be affected by the path. We can choose an appropriate value between 0 and 1 for r, according to the dependability of previous path data. For the rest handovers of the supposed path, the online learning process should continue to update $t_{b,c}^2$, $t_{b,j}^2$, $t_{c,d}^3$ and $t_{c,j}^3$ until the path data is over. At this time, the update of transition matrix for the current path data ends.

4) Step 4: As the update rule (8) of transition matrix would result in negative value for transition probability, the online learning process would amend transition matrix to keep every transition probability non-negative to satisfy the constraint (2). The method is to check the transition probability of the matrix row by row. The transition probabilities that are negative in each row are added and then averaged to other transition probabilities that are positive. These

transition probabilities with original negative value, are then set to zero.

Until there is no transition probability with negative value in the entire transition matrix, the online learning process ends and the next learning is ready.

B. The popularity of BS

The above scheme for user mobility prediction is only for a single UE. If the VNF resources for BS are dynamically allocated for a single UE, the signaling overhead for resource re-allocation will increase greatly. Hence, in order to improve the service capability of the network slice, we consider re-allocating VNF resources for BS that are subject to increased load resulting from the large-scale user handover. These BSs are defined as hotspot BSs (HBSs).

We let $\boldsymbol{H}=[H_1, H_2 \ldots H_n \ldots H_N]$ to denote the popularity of each BS in the network. The initial value for any element H_n is equal to zero. We assume that the online learning system have updated the transition matrix k times for each UE. Therefore, the k -step direct transition probability $\boldsymbol{p}_{k-m}=[\alpha_1^k \ \alpha_2^k \ \ldots \alpha_n^k \ \ldots \alpha_N^k]_m$ for UE m can be attained by

$$\boldsymbol{p}_{k-m}=\boldsymbol{p}_{0-m}*\boldsymbol{T}_{k-m}, \tag{9}$$

where $\boldsymbol{p}_{0-m}$ indicates the initial probability for the UE m, and $\boldsymbol{T}_{k-m}$ indicates the k -step transition matrix for the UE m.

Hence, the popularity of every BS can be calculated by

$$\boldsymbol{H}=\sum_{m=1}^{M}\boldsymbol{p}_{k-m}. \tag{10}$$

Apparently, if a BS has the high popularity, enormous UEs are likely to hand over to that BS.

IV. Dynamic Resource Allocation for VNF

In this section, we describe the main idea of dynamic resource allocation for VNFs on HBSs. Then we develop a heuristic algorithm to implement resource allocation for VNFs.

A. Resource allocation for VNF

As the VNFs are generally virtualized from some physical resources, we consider there are K types of resources, as exemplified by CPU, memory and bandwidth. The total amount of each type resource is denoted by Q^k. We further assume these resources have been virtualized to V types of VNFs. Denote by q_v^k the amount of type $k \in K$ resource for a type $v \in V$ VNF instance. Hence, denote by $R_n^{\circ k}$ the type k resource that has been allocated for VNFs on BS_n, which can be calculated by

$$R_n^{\circ k}=\sum_{v=1}^{V} q_v^k x_n^v, \ \forall k \in K, \tag{11}$$

where x_n^v represents the number of type v VNF instances that have been deployed on BS_n.

Due to the user mobility, the allocated resource on BSs generally cannot respond to the change of service load. However, with the help of the popularity of BSs, the resource is capable of being allocated dynamically to the HBSs. We let R_n^{*k} to denote the expected amount of type k resource that should be allocated for VNFs on BS_n. As the intuition that the most resource should be allocated to where the most demands exist, we consider that R_n^{*k} should complies with the following

$$R_n^{*k}=\frac{H_n}{\sum_{i=1}^{N} H_n}*Q^k. \tag{12}$$

Moreover, we further define the allocation tolerance G_n^k as

$$G_n^k=R_n^{\circ k}-R_n^{*k}. \tag{13}$$

Hence, in order to guarantee the QoS of UEs that execute handover, the efficient resource allocation for VNFs on BSs can be formulated as

$$\min_{x_n^v} \sum_{n=1}^{N}\sum_{k=1}^{K}\left|G_n^k\right|, \text{ s.t. } \sum_{n=1}^{N} q_v^k x_n^v \le Q^k. \tag{14}$$

B. Complementarity Mechanism (CM)

Although the objective function of resource allocation for BS can be treated as a least-square problem, the problem is still difficult to solve as the optimization variable x_n^v is an integer. Hence, we propose a complementarity mechanism. The main idea of the complementarity mechanism is to allocate idle resources of those base stations with fewer handover users to HBSs, instead of allocating resource directly from the total amount of resource. Specifically, the mechanism reduces the excessive signaling cost and avoids recalculating the amount of resources x_n^v that need to be allocated to the VNFs on HBSs. The key procedures of mechanism implementation can be summarized as the following:

1) Step 1: We first initialize the system parameters, such as the BS distribution, the user movement path data for prediction, and the deployment of VNFs on each BS. Meanwhile, considering that resource scheduling will bring additional signaling overhead to the system, we set the maximum number of resource compensation times D for the mechanism.
2) Step 2: We perform the online learning process, and attain the predicted result on group user mobility. Then we calculate the popularity of each BS according to formula (10).
3) Step 3: We acquire the initialized resource allocation of VNFs on each BS to represent $R_n^{\circ k}$ (i.e., the actual allocation). Meanwhile, the ideal resource allocation R_n^{*k} (i.e., the objective allocation) of VNFs on each BS is calculated by formula (12).
4) Step 4: Then we calculate the difference G_n^k between the actual resource allocation and the ideal resource allocation for each BS.
5) Step 5: According to the value of G_n^k, we divide all BSs into three categories, the BS with excess demand (CDBSs) (i.e., $G_n^k > 0$), the BS with insufficient demand (IDBSs) (i.e., $G_n^k < 0$), and the BS with matching demand (MDBSs) (i.e., $G_n^k = 0$).

6) Step 6: We first sort the IDBSs according to the popularity of each IDBS. In the descending order, for each type of resource on each IDBS, we select the appropriate CDBS to be paired with each IDBS (i.e., $\left|G_i^k\right|_{\text{IDBS}} \le \left|G_j^k\right|_{\text{CDBS}}$), and schedule the excess resource from paired CDBS to the corresponding IDBS. Based on the idea of "returning more and less compensation", we give priority to ensuring the demand of IDBS is satisfied fully (i.e., $\left|G_i^k\right|'_{\text{IDBS}} \ge 0$, where $\left|G_i^k\right|'_{\text{IDBS}}$ represents the allocation tolerance after the compensation). Specifically, we call one time compensation as the completion of resource scheduling for all types of resource on specific IDBS. As being taken away the excess resource, CDBS would become an IDBS or MDBS. Hence, after one time compensation, we update the set of IDBSs, the set of CDBSs, and the set of MDBSs.
7) Step 7: Due to the limit of available resource and the minimum granularity of resource scheduling for each VNF, all the demands of IDBSs may not be satisfied fully. Hence, we design two termination criteria for the mechanism. One is to terminate the compensation when there is no IDBS in the system. The other is to terminate the compensation when the times of compensation achieve the maximum number of compensation times D .

The following heuristic algorithm is developed to implement the complementarity mechanism.

Algorithm: (Complementarity Mechanism)

Input: BS distribution, user movement path data, deployment of VNFs on each BS (i.e., x_n^v), parameter D

Output: the actual resource allocation $x_n'^v$ after the resource compensation

```
Initialization
Predict the user mobility by formula (9)
Calculate the popularity of each BS by formula (10)
Calculate R^{v,k}_n by formula (11) based on the deployment of VNFs
Calculate R^{*k}_n by formula (12)
Calculate G^k_n by formula (13)
Create {IDBS}, {CDBS}, {MDBS} based on G^k_n
Sort {IDBS} based on the popularity of each IDBS
for d=1 to D do
  for i=1 to length({DBS}) do
    for k=1 to K do
      if |G^k_i|_IDBS ≤ |G^k_j|_CDBS (pairing)
        x'^v_n = x^v_n + round^-(|G^k_j|_CDBS − |G^k_i|_IDBS)
      end if
    end for
    d=d+1
    Update {IDBS}, {CDBS}, {MDBS}
  if d < D
    continue;
  else
    break
  end for
  break
end for
```

In the algorithm, *length* (.) represents the function to calculate the number of elements in a set, and *round*$^-$ (.) represents the function to make the value round down.

V. Performance Evaluation

In this section, we compare Complementarity Mechanism (CM) with the other two resource allocation schemes: fairness allocation (FA) scheme and random allocation (RA) scheme. The FA scheme allocates the resource fairly for each type of VNF on each BS. The RA scheme allocates the resource randomly.

A. Simulation Settings

We consider four scenarios for simulation. As shown in Table I, different key parameter changes in each scenario. The simulation parameters are UEs number, handover number of each movement path for each UE, the types of resource, the amount of each type resource, the types of VNF, and the maximum number of compensation times. The typical value for each parameter references by [14], and the movement path data come from [13]. Specifically, In order to simplify the illustration of resource utilization, we further define the resource utilization for one type resource as $\eta^k = (Q^k - \sum_{n=1}^{N}\left|G_n^k\right|)/Q^k$, then the final resource utilization is denoted by $\sum_{k=1}^{K}\eta^k / K$.

TABLE I. Simulation Parameters

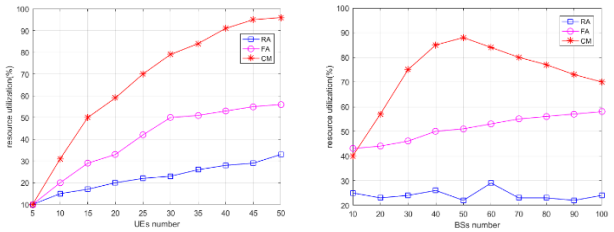
Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
UEs number	variable	30	30	30
handover number	10	10	variable	10
BSs number	37	variable	37	37
the types of resource	3	3	3	3
resource 1# amount	10	10	10	10
resource 2# amount	20	20	20	20
resource 3# amount	50	50	50	50
the types of VNF	5	5	5	5
compensation Max.	10	10	10	variable

B. Numerical Results

In experiment 1, we investigate the resource utilization of every resource allocation scheme when UEs number changes. As shown in Fig. 3 (a), the proposed CM is obviously superior to FA and RA. Meanwhile, with the increase of UEs number in the system, the trends of the three schemes are different. Since more UEs bring more movement path data, online learning system is capable of predicting user mobility in higher accuracy. Hence, the resource utilization for CM increases as HBSs permit more handover users. For the FA, more UEs also lead to the increased resource that are allocated fairly to each BS to provide service for new accessed users. However, the fairness of resource allocation in the FA cannot guarantee the resource of each BS is utilized fully, as some BS may permit a few users. Due to the randomness of resource allocation, the RA may allocate a large amount of resources to some BSs that do not accept new access users, so the resource utilization is the lowest.

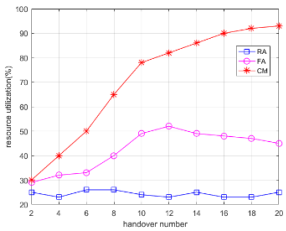
In experiment 2, we investigate the influence of BSs number on resource utilization. As shown in Fig. 3 (b), with the increase of BSs number, the CM is better than FA and RA. As the BSs number increases, the resource utilization for CM increases first, and then reaches a maximum value about

88% and then decreases. The main reason is that the transition probability of CM is hard to converge as there are too many BSs. Namely, the popularity of each BS cannot be distinguished, and each BS has similar popularity. The CM has to allocate resource equally to each BS. Therefore, the resource utilization for CM becomes similar to the FA.

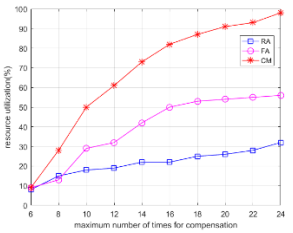


(a) Scenario 1

(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

Fig. 3. Comparison of the resource utilization for CM, FA, and RA in different scenarios

In experiment 3, we investigate the influence of handover number on resource utilization. As shown in Fig. 3 (c), with the increase of handover number, the resource utilization for CM is significantly better than FA and RA. Moreover, when the handover number continues to increase, the resource utilization for CM can arise up to 90%, while the other two schemes show a downward trend.

In experiment 4, we investigate the influence of the maximum number of compensation times on resource utilization. As shown in Fig. 3 (d), as the maximum number increases, the resource can be scheduled richly among IDBSs and CDBSs, which would lead the allocation tolerance to be zero. Hence, the resource utilization for CM is capable of achieving 100% nearly. As there is no limit on the number of resource allocations in the FA and the RA, the maximum number of compensation times has no effect on their resource utilization.

VI. Conclusion

In this paper, we study the resource allocation for VNFs on each BS in slice based 5G network. In order to guarantee the QoS of mobile user when handover occurs, we propose a complementarity mechanism to allocate resource to HBSs that would permit a large number of new accessed users. As employing online learning system to predict the group user mobility, we can identify the HBSs easily, and schedule the idle resource from other BSs to the HBSs. Numerical result validate the effectiveness of our proposed complementarity mechanism.

Acknowledgement

This work is supported in part by the National Science Foundation of China (No. 61471089), Scientific Research Fund of Sichuan Provincial Education Department (No. 17ZB0355), and Open Research Fund of Chongqing E-commerce and Modern Logistics Laboratory (No. ECML01710).

References

[1] A. Osseiran, F. Boccardi, V. Braun, and K. Kusume, et al, “Scenarios for 5G mobile and wireless communications: the vision of the METIS project,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26-35, May 2014.

[2] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, “NFV: state of the art, challenges, and implementation in next generation mobile networks ,” *IEEE Netw.*, vol. 28, no. 6, pp. 18-26, Nov. 2014.

[3] B. Cao, F. He, Y. Li, and C. Wang, et al, “Software defined virtual wireless network: framework and challenges, ” *IEEE Netw.*, vol. 29, no. 4, pp. 6-12, 2015.

[4] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, “Resource slicing in virtual networks: a survey,” *IEEE Trans. on Netw. and Serv. Manag.*, vol. 13, no. 3, pp. 462-476, 2016.

[5] X. An, C. Zhou, R. Trivisonno, and A Kaloxylos, et al, “On end to end network slicing for 5G communication systems,” *Trans. on Emerg. Tele. Tech.*, vol. 28, no. 4, e3058, 2017.

[6] I. D. Silva, G. Mildh, A. Kaloxylos, and P. Spapis, et al. “Impact of network slicing on 5G radio access networks,” Networks and Communications (*EuCNC*), 2016 European Conference on. IEEE, 2016, pp. 153-157.

[7] Y. Zaki, L. Zhao, C. Goerg. and A. Timm-Giel, “LTE wireless virtualization and spectrum management,” in Proc. 3rd Joint IFIP Wireless Mobile Netw. Conf. (*WMNC*), Budapest, Hungary, 2010, pp. 1-6

[8] Y. Zaki, L. Zhao, C. Goerg. and A. Timm-Giel, “LTE mobile network virtualization,” *Mobile Netw. Appl.*, vol. 16, no. 4, pp. 424-432, Jun. 2011.

[9] M. I. Kamel, L. B. Le, and A. Girard. “LTE wireless network virtualization: dynamic slicing via flexible scheduling,” Vehicular Technology Conference (*VTC Fall*), IEEE, 2014, pp. 1-5.

[10] E. Pateromichelakis, and C. Peng. “Selection and dimensioning of slice-based RAN controller for adaptive radio resource management,” IEEE Wireless Communications and Networking Conference (*WCNC*), 2017, pp. 1-6.

[11] K. Zhu and E. Hossain, “Virtualization of 5G celluar networks as a hierarchical combinatorial auction,” *IEEE Trans. Mobile Computing*, vol. 15, no. 10, pp. 2640-2654, Oct. 2016.

[12] A. Mohamed, O. Onireti, S. A. Hoseinitabatabaei, and M. Imran, et al,“Mobility prediction for handover management in cellular networks with control/data separation,” IEEE International Conference on Communications (*ICC*), 2015: 3939-3944.

[13] SpazioDati. Dandelion API[EB/OL].[2017-12-11.]. https://dandelio-n.eu/datamine/open-big-data/.

[14] M. F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutabe. “On orchestrating virtual network functions,” Network and Service Management (*CNSM*), 11th International Conference on. IEEE, 2015: 50-56.