Joint Resource Allocation and Online Virtual Network Embedding for 5G Networks

Junling Li, Ning Zhang, Qiang Ye, Weisen Shi, Weihua Zhuang, and Xuemin (Sherman) Shen
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, N2L 3G1
Email:{j742li, n35zhang, q6ye, w46shi, wzhuang, sshen}@uwaterloo.ca

***Abstract*—Next generation (5G) wireless networks are expected to accommodate proliferation of connected devices and multimedia services. To support multimedia services in an agile, cost-effective, and flexible way, network virtualization is a potential solution. This paper investigates service-oriented network virtualization for 5G wireless networks, to efficiently allocate heterogeneous resources to accommodate multimedia services. Specifically, we study joint resource allocation for virtual network requests (VNRs) and online embedding the resultant VNRs in core networks (CNs). With the deployment of multiple traffic aggregation points (TAPs) in radio access networks (RANs), the end-to-end traffic from heterogeneous access technologies can be aggregated and then grouped based on their destinations. Queueing models are developed in determining the minimal capacity required at each core network element. Virtual network embedding (VNE) in the core network is further proposed to achieve efficient physical resource sharing in CNs. Simulation results validate the VNE process in core networks based on the optimized capacities.**

***Index Terms*—5G, network virtualization, virtual network embedding (VNE), resource allocation.**

I. Introduction

With explosive development of mobile Internet and Internet of things (IoT), emerging services and applications with diverse requirements and characteristics will prominently arise in future 5G networks [1]-[3]. Various application scenarios will have different quality-of-service (QoS) requirements in terms of latency, security level and complexity [4]. For example, the vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications in vehicular ad-hoc networks (VANETs) feature high mobility, e-health services require high security, Augmented Reality (AR) requires ultra-low latency, and wearable equipment needs low complexity and high reliability. However, the current one-type-fits-all mobile networks cannot provide users with customized services well [5].

To accommodate the aforementioned differentiated services, heterogeneous resources should be intelligently integrated into 5G networks with efficient allocation [6]. The physical resource pool manifests high heterogeneity, ranging from radio access networks (RANs) to core networks (CNs). For the RAN segment, spectrum and time slots are the major resources that can be allocated by operators, while for the CN segment, the resources include computing, communication and storage resources at each CN element (nodes and links). In addition, to support massive connected intelligent terminals and provide seamless connectivity, macro-cells, small cells, and femtocells are expected to be densely deployed, which composite heterogeneous multi-tier access networks in future 5G networks [7]. Various wireless access technologies (e.g. LTE, Wi-Fi, WCDMA, etc.) will coexist and cooperate with each other to better utilize network resources. Through efficient integration, the service quality is expected to be enhanced [8].

The heterogeneity in services, resources, and access technologies increases the complexity of 5G networks, leading to high costs of deployment and maintenance. Thus, the development of 5G networks is not simply an inheritance from the current 3GPP mobile network, but an evolution of the network architecture [9]. A scalable, delay-optimal, highly adaptable and flexible network architecture is desired, which has attracted more and more attentions from both the industry and the academia.

Network virtualization [10][11], or network slicing, has great potential to address the above challenges. Through network virtualization, the whole network is sliced into multiple virtual networks (VNs), and each VN is used to support a customized service in a specific scenario. The heterogeneous resources are virtualized to allow resource sharing among different VNs and to improve resource utilization in both RANs and CNs. As a result, the customized services can be accommodated in a cost-effective, agile and flexible manner.

In last few years, many research activities have been carried out to implement network virtualization in different scenarios, ranging from wired networks [10] to wireless networks [12]. Network virtualization decouples the role of current service providers into two new roles: the Infrastructure Providers (InPs) which deploy and maintain the substrate network equipment and Virtual Network Operators (VNOs) which assemble, install and manage virtual networks and provision customized services to users. The problem of mapping multiple virtual networks onto a given substrate network is the major task in network virtualization and usually termed as Virtual Network Embedding (VNE). However, most of the existing VNE approaches consider to map virtual networks with predefined constraints onto a physical substrate network. To our best knowledge, there is no work that considers optimizing each VN's resource requirements and VNE problems in a combined way.

In this paper, we investigate network virtualization for 5G networks to efficiently allocate heterogeneous resources to accommodate multimedia services. A novel framework that jointly considers the optimization of the virtual network resource requirement and the VNE problem in the CN is

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

presented. The framework consists of two major stages. In the first stage, the optimization goal is to minimize the leasing costs while satisfying the given end-to-end QoS requirements. Traffic from different access networks are first aggregated and then used to find the optimal resource capacity for each VN based on M/M/1 queueing model. In the second stage, multiple VNs are formed by virtual network operators, each representing a customized end-to-end service with the capacity constraints from the previous stage. Through mapping these VNs onto the CN elements using efficient VNE algorithms, multiple VNs can share the common CN resource pool to improve the physical resource utilization.

The remainder of this paper is organized as follows. The system model is provided in Section II. The optimization of VN resource requirement is presented in Section III. In section IV, simulation results are provided to validate the processes of determining the optimal capacities and the VNE in the core network. Section V concludes the paper.

II. System Model

A. Service-Oriented Network Virtualization in 5G HetNets

The HetNet environment considered in this paper is illustrated in Fig. 1. We consider a scenario where different access technologies, (e.g. LTE, Wi-Fi, DSRC, etc.) coexist in a RAN. The wireless access points (WAPs) are deployed to support massive communications in the environment. We design a novel framework for service-oriented core network virtualization in the HetNet. The framework consists of four parts, i.e. service priority assignment, traffic aggregation in the RAN, logical topology generation and virtual network embedding in the CN.

1) Service Priority Assignment: Due to the different QoS requirements in 5G networks, we classify the services into three main categories in terms of the latency requirement, i.e., ultra-low delay, low delay, and delay tolerant services. In this work, we assign the highest priority for services with ultra-low delay requirement (less than 20 ms), such as augmented reality, and e-health services. The medium priority is assigned to the services with relatively low delay requirement, such as voice over IP (VoIP) and video conferencing. Some services such as file transfer and e-mail require 100% reliable data transfer but with lower priority for latency, so we assign the lowest priority for them.

2) Traffic Aggregation in the RAN: We consider a scenario where multiple traffic aggregation points (TAPs) [13] are deployed in the RAN. Each TAP has the capability of aggregating traffic from its adjacent WAPs and is aware of other TAPs' locations. We can consider several adjacent WAPs within a certain area as a *Cluster*, and one TAP that connects to all these WAPs as a *Cluster Head*. Then, the *Cluster Head* is used to aggregate all the end-to-end communication requests generated by the cluster.

3) Logical Topology Generation: After traffic aggregation, each TAP groups traffic into subgroups based on their destinations. Service requests with the same TAP as destination will be grouped together as a subgroup. Each subgroup may contain service requests with all three priorities. With the statistics of service requests and queueing theory modeling, multiple end-to-end logical topologies with optimal node and link capacity requirements are formed by VNOs. The optimal capacities can be determined by minimizing the cost for leasing physical resources from the CN while satisfying the average QoS (e.g. delay) requirements for services with different priority levels.

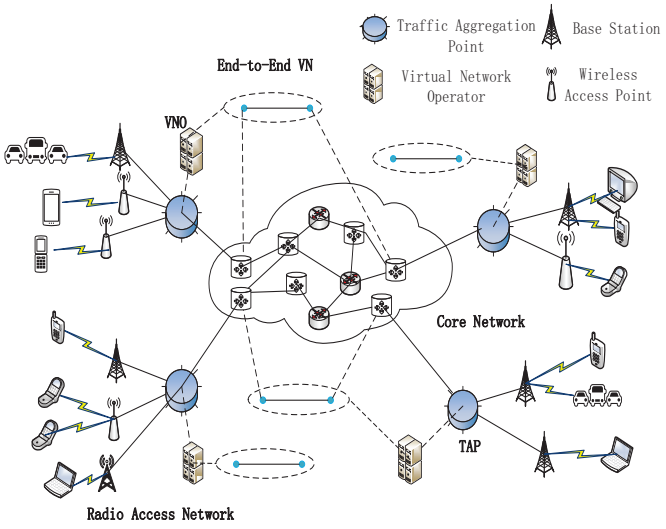


Fig. 1. Service-oriented core network virtualization in HetNets

4) Virtual Network Embedding in the CN: Given the logical topologies created by multiple VNOs, finding an optimal logical-physical resource mapping is referred to as the VNE problem. Each of these logical topologies is called an end-to-end virtual network request (VNR). The mapping strategy needs to consider the capacity requirements of virtual nodes and virtual links as well as the physical resource constraints. The performance of an embedding algorithm can be measured using various metrics, such as the long-term average revenue [14], the average embedding cost per VNR [15], and the acceptance ratio. Through efficient logical-physical resource mapping, multiple VN requests can coexist on the common physical infrastructure. By resource sharing, the improvement on the utilization of CN elements can be achieved. In addition, the management of physical resources becomes more flexible as VNRs are isolated and independent with each other, and each VN can be managed by its own VNO separately.

B. Determining Optimal Capacity via Queueing Model

Let M represent the total number of TAPs in the network. From the i-th TAP, we model the arrival of service requests as a Poisson process with arrival rate $\lambda_i, i = 1, 2, ..., M$. After traffic grouping, the stream is grouped into M sub-streams, with the arrival rate of each sub-stream being $\lambda_{ij}, j = 1, 2, ..., M$. Each sub-stream can be further classified into three subsets with arrival rate $\lambda_{ijk}, k = 1, 2, 3$, based on the service priority levels. Then, we have

$$\lambda_i = \sum_{j=1}^{M} \sum_{k=1}^{3} \lambda_{ijk}, \tag{1}$$

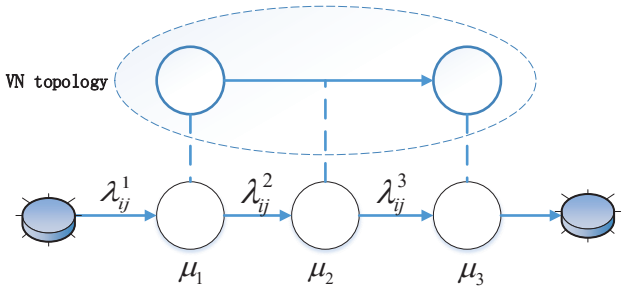


Fig. 2. The queueing network to determine the optimal capacity

$$\lambda_{ij} = \sum_{k=1}^{3} \lambda_{ijk}. \tag{2}$$

Now we only consider λ_{ij}, one of the sub-streams coming out of TAP i whose destination is TAP j. The VNO needs to form an end-to-end VN topology for λ_{ij}. As illustrated in Fig. 2, an end-to-end VN topology consists of two virtual nodes and one virtual link. Each virtual element can be treated as a serving point. For each subset λ_{ijk} in λ_{ij} whose service priority level is k, we consider it as the traffic source of a queueing network with three M/M/1 queueing systems. Each M/M/1 queue is a queue where the arrival process is Poisson and the service times are exponentially distributed. The first-come first-served (FCFS) queueing discipline is adopted.

Let each serving point be with independent service rate $\mu_n, n = 1, 2, 3$, giving service times that are exponentially distributed with mean $1/\mu_n$. Since there are three types of services supported by each serving point, we consider μ_n as the sum of $\mu_{nk}, k = 1, 2, 3$, where μ_{nk} represents the service rate of the n-th serving point to support the service with priority k. For the n-th queueing system, the Poisson arrival process is with mean arrival rate λ_{ij}^n. We consider all the substrate nodes as switches (without traffic inflation and deflation), therefore,

$$\lambda_{ij} = \lambda_{ij}^n, n = 1, 2, 3. \tag{3}$$

λ_{ij}^n can be divided into three subsets λ_{ijk}^n based on the service priority. The traffic intensity for the priority k, ρ_{nk}, is given by

$$\rho_{nk} = \lambda_{ijk}^n / \mu_{nk}. \tag{4}$$

For stability, it requires that $\rho_{nk} < 1$ for the queue to be at equilibrium. For a FCFS M/M/1 queue at equilibrium, we define T_{nk} as the waiting time in the system and M_{nk} as the number of customers in the system. Then, the expectation of T_{nk} can by given by

$$E[T_{nk}] = \frac{1}{\mu_{nk} - \lambda_{ijk}^n}. \tag{5}$$

According the Little's law, we have the average number of customers in the system as follows:

$$E[M_{nk}] = \lambda_{ijk}^n E[T_{nk}] = \frac{\lambda_{ijk}^n}{\mu_{nk} - \lambda_{ijk}^n}. \tag{6}$$

We consider the latency requirement (i.e., time for a packet to pass from one TAP to another TAP) for the whole system and show how to meet the latency requirement of services with different priorities with minimum service rate at each element. Suppose that the latency is required to be less than T_k, for the services with priority level k, $k = 1, 2, 3$. According to the Little's law, we have the following constraint

$$\sum_{n=1}^{3} E[M_{nk}] = \sum_{n=1}^{3} \frac{\lambda_{ijk}^n}{\mu_{nk} - \lambda_{ijk}^n} \leq \lambda_{ijk} T_k. \tag{7}$$

Let the unit price vector of service rates at CN elements be $\boldsymbol{p} = \{p_1, p_2, p_3\}$, and then the objective of each VNO is to minimize the total cost $C(\mu_1, \mu_2, \mu_3)$, where

$$\begin{aligned} C(\mu_1, \mu_2, \mu_3) &= \sum_{n=1}^{3} p_n \mu_n, \\ &= \sum_{n=1}^{3} p_n (\sum_{k=1}^{3} \mu_{nk}). \end{aligned} \tag{8}$$

The following constrained optimization problem can be formulated to determine the optimal capacities of each VNR

$$\begin{aligned} (\mu_1^*, \mu_2^*, \mu_3^*) = \arg \min_{(\mu_1, \mu_2, \mu_3)} & C(\mu_1, \mu_2, \mu_3) \\ \text{s.t.} \ \sum_{n=1}^{3} \frac{\lambda_{ijk}^n}{\mu_{nk} - \lambda_{ijk}^n} &\leq \lambda_{ijk} T_k, \\ \mu_n &\in (\lambda_{ij}, \mu_{n,max}], \end{aligned} \tag{9}$$

where $\mu_{n,max}$ represents the maximum capacity of the n-th CN element. The objective of this constrained optimization problem is to minimize a linear function with several linear inequality constraints and bound constraints. By solving it, the optimal processing capacity at each CN element $(\mu_1^*, \mu_2^*, \mu_3^*)$ can be obtained.

III. Online Virtual Network Embedding in Core Networks

A. Problem Formulation

Formally, the VNE problem in CNs can be described as follows:

Let the substrate network be denoted by an undirected graph, $G^s = (N^s, L^s, \mu_N^s, \mu_L^s)$, where N^s and L^s represent the substrate nodes and substrate links, respectively. Suppose that each node and link in the physical substrate network is capable of storing, processing and switching data packets, but with various capacities. In other words, all the substrates nodes and links are associated with service rates as their attributes, denoted by μ_N^s and μ_L^s, respectively. The set of all loop-free paths in the substrate network is denoted by P^s. One example of a substrate network is shown in the right side of Fig. 3. The numbers on the links represent available bandwidth while the numbers in the rectangles indicate the available CPU resources of each node.

Let the virtual network request be represented by $G^v = (N^v, L^v, \mu_1, \mu_2, \mu_3, P_l)$, where N^v, L^v, and P_l represent the virtual nodes, the virtual link, and the priority level of the VN

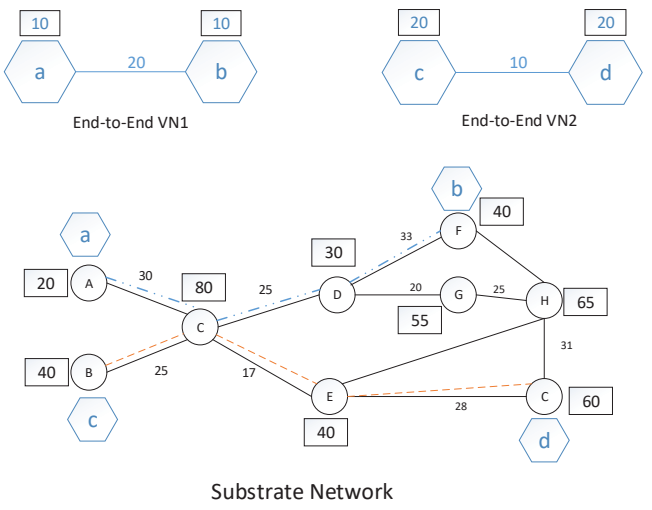


Fig. 3. Illustration of the VNE problem

request, respectively. In this study, all the VN requests represent end-to-end communications, and thus each VN consists of two virtual nodes and one virtual link connecting them. A VN request has two node capacity constraints (μ_1 and μ_3) and one link capacity constraint (μ_2) specified in terms of the substrate attributes with it. Two VN requests are shown on the top of Fig. 3: VNR1 requires bandwidth 20 over the link (a, b) and CPU capacity 10 at each node, while VNR2 requires bandwidth 10 over the link (c, d) and CPU capacity 20 at each node. For a given VN request G^v, the definition of the VNE problem is to find a mapping M that embeds G^v onto a subset of G^s while satisfying the constraints in G^v, i.e.,

$$M : G^v \rightarrow (N^{alloc}, P^{alloc}, R_N, R_L),$$

where $N^{alloc} \subset N^s$ and $P^{alloc} \subset P^s$, R_N and R_L are the substrate (node and link) resources allocated to the VN requests. The VN embedding procedure can be divided into node-mapping M^N and link-mapping M^L as follows:

$$M^N : (N^v, \mu_1, \mu_3) \rightarrow (N^{alloc}, R_N),$$
$$M^L : (L^v, \mu_2) \rightarrow (P^{alloc}, R_L).$$

It has been demonstrated that solving the VNE problem is NP-hard [16]. Directly solving the VNE problem to obtain the truly optimal solution is feasible only for small substrate sizes. Therefore, heuristic or meta-heuristic approaches are the focus of the literature to find near optimal solutions.

The VNE process in our work can be regarded as the post process of the first stage in the proposed framework. In fact, any existing VNE approaches developed for wired networks can be applied. In this study, we choose two basic yet widely used heuristic approaches to do online VNE in the core work. We argue that the current VNE algorithms are just used for demonstration purpose, and more advanced algorithms are to be developed and applied in the future work.

B. Handling Online VN Requests

The ability of handling online VN requests of our VNE algorithm can be illustrated at two levels. At the first level,

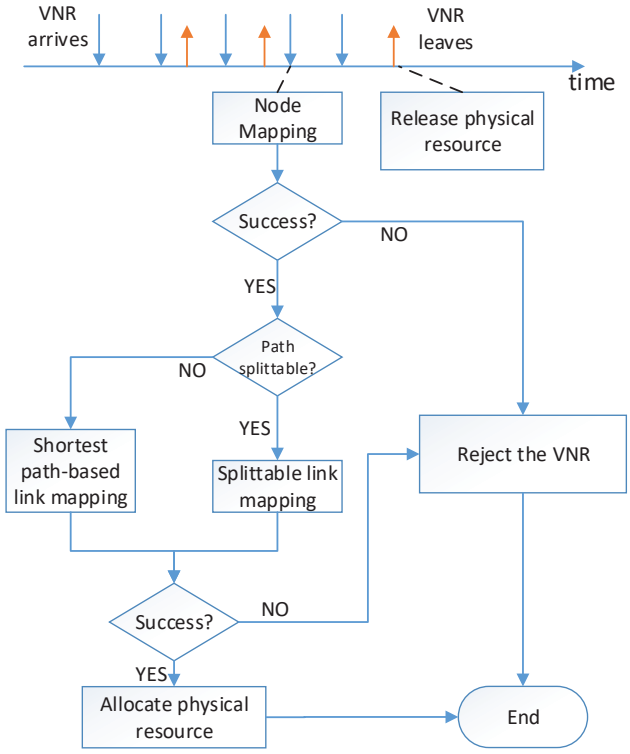


Fig. 4. Online virtual network embedding procedure

the traffic aggregation is performed periodically at each TAP. Each traffic flow coming out of TAP i has its lifetime $t(\lambda_{ij})$. Each time when a TAP finishes aggregating the traffic from its nearby WAPs, it will wait until all the created VNRs expire to start the next round of traffic aggregation.

At the second level, the online VNE problem for all the TAPs are considered. The new VN requests created by all the TAPs arrive and old VN requests leave the substrate continually over time. Let the during time (lifetime) of a VN request representing end-to-end communication between TAP i and TAP j be $t^v_{i,j}$, then

$$t^v_{i,j} = t(\lambda_{ij}). \tag{10}$$

VNE is performed at each time a new VNR arrives by performing the node and link mapping algorithms. If a successful mapping is found for the VNR, the corresponding physical resource will be allocated accordingly until its lifetime expires. Otherwise, the VNR will be rejected. If a VNR lifetime expires, the resource occupied by the VNR would be released. Fig. 4 illustrates the overall online VNE procedure in the core network.

C. Node Mapping Algorithm

A greedy node mapping algorithm is employed as the first phase of the VNE procedure. The basic idea is to map the virtual nodes with more constraints to the substrate nodes with more residual node resources so as to minimize the use of the resources at the bottleneck nodes/links.

When a new VN request arrives, a subgroup of substrate nodes N^s_{sub} that can meet the node constraints (CPU capacity)

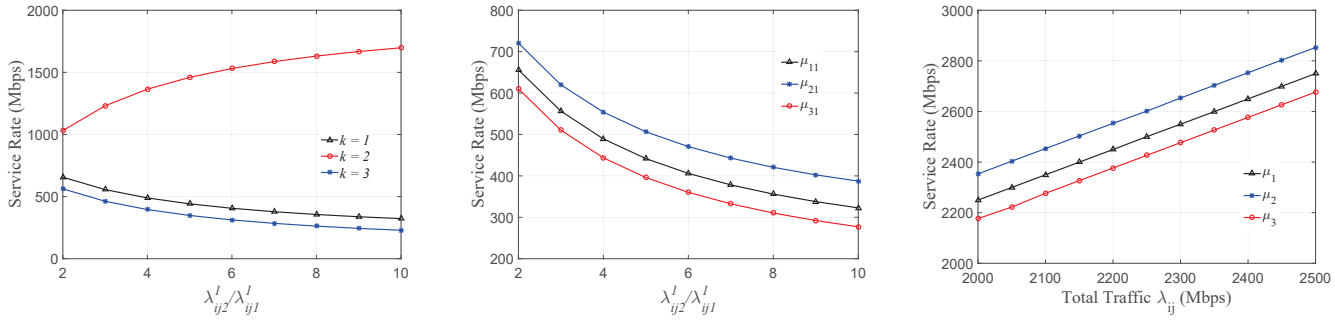


Fig. 5. Optimal capacity determination results with varying traffic composition and data volume (a) optimal capacity needed for different types of services at the first CN element (b) optimal capacity needed at three CN elements for the services with highest priority c) optimal capacities needed at three CN elements for different arrival rates

are first selected as candidates for each virtual node. If N^s_{sub} is empty, the VNR would be rejected. Otherwise, N^s_{sub} will be sorted in descending order according to their CPU capacities. Then, the virtual node is mapped to the substrate node with highest CPU capacity in N^s_{sub}.

D. Link Mapping Algorithms

We employ two link mapping algorithms: shortest path-based link mapping algorithm [17] and splittable link mapping using MCF [18]. We assume that path splitting is supported by the substrate network. Link mapping algorithm is determined depending on the VNR's splitting choice. For shortest path-based link mapping, we aim to find the k-shortest paths for increasing k for each virtual link of the VNR. If a path with enough bandwidth capacity is found, the virtual link is mapped to that path. Otherwise, the VNR would be rejected. For the splittable link mapping algorithm, the virtual link mapping is completed by applying the multi-commodity flow algorithm. In this work, we will implement both algorithms and compare their performance in terms of different metrics.

IV. Performance Evaluation

A. Optimal Capacity Determination Results

We present a case study to demonstrate how VNOs determine the optimal resources required to support the respective services. Let the unit price vector of service rates at substrate elements be $(p_1, p_2, p_3) = (0.1, 0.05, 0.2)$. The delay requirements of services with highest, medium, and lowest priority are 20ms, 50ms and 100ms, respectively. Assuming that the average arrival rate λ_{ij} of the total traffic is 2000 Mbps. In reality the constitutional proportion of λ_{ij}, i.e. $\lambda_{ij1} : \lambda_{ij2} : \lambda_{ij3}$ is affected by multiple factors (e.g. TAP location, user density, time etc.). Since the data volume of the second kind services is typically much higher than the other two kinds of services, we consider a simple case where $\lambda_{ij1} = \lambda_{ij3}$, and $\lambda_{ij1} : \lambda_{ij2}$ ranges from 1:10 to 1:2. The maximum service rate at each substrate element is set to be 3000 Mbps.

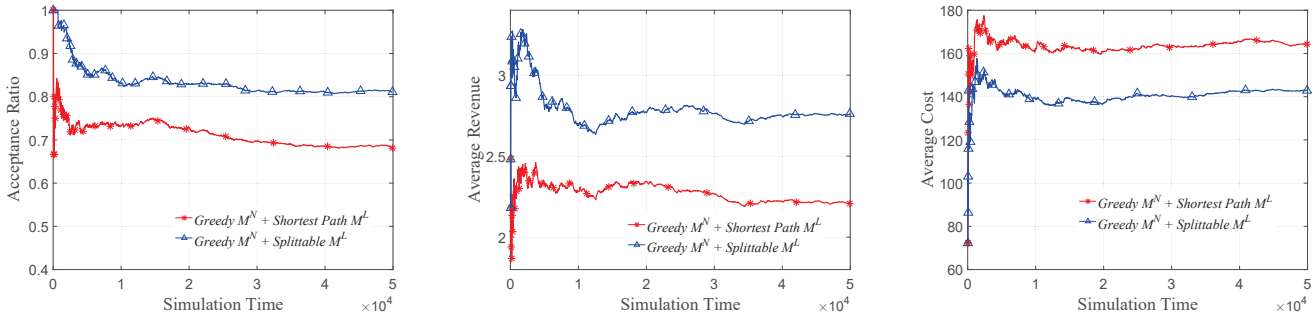
Using the parameters above, the constrained optimization problem (9) is solved using the constrained optimizer in MATLAB. Fig. 5(a) shows the optimal capacity needed at the first CN element ($n = 1$) for three kinds of services with different constitution proportion of λ_{ij}. It can be seen that as the proportion of lowest priority services increases, the capacity allocated to it goes higher while the capacity allocated to the other two services goes down. Next, we show the optimal capacity needed at three CN elements for the services with the highest priority ($k = 1$) in Fig. 5(b). As can be seen from the figure, the capacities at all the three elements become smaller as the arrival rate of the service decreases. To minimize the total leasing cost, the higher unit price of leasing an element is, the lower capacity is assigned to that element. Lastly, we vary the arrival rate λ_{ij} in the range of [2000, 2500] (Mbps) and set the $\lambda_{ij1} : \lambda_{ij3} : \lambda_{ij2}$ to be fixed at 1:1:5. Fig. 5(c) shows the optimal capacities required at each CN element when the total traffic varies.

TABLE I
Simulation Settings for VNE in CN

Parameter	Value	Parameter	Value
number of nodes	50	VNR arrival rate	25
number of links	130	total simulation time	50000
service rate at each CN element	[2500, 3500]	arrival rate of service requests	[1500, 2500]

B. Virtual Network Embedding Results

In this section, we show the virtual network embedding results in the core network. The substrate network is randomly generated with 50 nodes using the GT-ITM tool [19] in a (50×50) grid. Each pair of substrate nodes is randomly connected with probability 0.5. The service rates of the substrate nodes and links are real numbers uniformly distributed between 2500 Mbps and 3500 Mbps. Each substrate node is connected to at least one TAP in the RAN. We assume that the VN requests generated by the VNOs arrive in a Poisson process with average arrival rate of 4 VNs per 100 time units. Each VN request has an exponentially distributed lifetime with an average of 1000 time units. For each TAP, we assume the arrival rate of service requests λ_{ij} is uniformly distributed between 1500 Mbps and 2500 Mbps. We consider a scenario



(a) (b) (c)

Fig. 6. Performance comparison between the two link mapping algorithms in terms of different metrics (a) acceptance ratio with time varying (b) average revenue with time varying (c) average cost per VNR with time varying

where $\lambda_{ij1} = \lambda_{ij3}$, and $\lambda_{ij1} : \lambda_{ij2}$ is a random value in the range of [0.1, 0.5]. Table I summarizes the simulation settings for the VNE validation process in the core network.

To compare the two link mapping algorithms, we first assume all the VNRs choose shortest path as their link mapping algorithm, and then try splittable link mapping for the same set of VNRs. For both cases we evaluate the acceptance ratio, average revenue and average cost with time varying. Fig. 6 shows the VNE results for the two cases. From Fig. 6(a) we can see that with splittable link mapping, the number of accepted VN requests is increased by about 15%. From Fig. 6(b) and Fig. 6(c) we can observe that splittable link mapping leads to higher average revenue and lower average cost for the core network.

V. Conclusion and Future Work

In this paper, we have proposed a novel framework that jointly considers optimal capacity determination for VNRs and physical resource sharing among the resultant VNRs, to efficiently support multimedia services in 5G networks. We have illustrated that the optimal capacity at each CN element can be determined via queueing modeling. Then the resultant virtual networks have been mapped to physical substrate network using heuristic VNE algorithms, promoting efficient resource sharing among multiple virtual networks.

For the future work, we will take service interruption into consideration in modeling the departure process. Moreover, we will develop more advanced service priority assignment schemes to accommodate highly diversified services.

Acknowledgment

This work is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] Q. Ye, and W. Zhuang, "Distributed and adaptive medium access control for Internet-of-Things enabled mobile networks," *IEEE Internet of Things Journal* vol. 4, no. 2, pp. 446-460, 2017.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065-1082, 2014.

[3] The 5G Infrastructure Public Private Partnership, "5G Vision: The 5G Infrastructure Public Private Partnership: The Next Generation of Communication Networks and Services," Feb. 2015.

[4] A. Osseiran, Boccardi *et al.*, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26-35, 2014.

[5] N. Zhang, P. Yang, S. Zhang, D. Chen, W. Zhuang, B. Liang, and X. Shen, "Software defined networking enabled wireless network virtualization: challenges and solutions", *IEEE Network* 2017.

[6] H. Li, K. Ota, and M. Dong, "Network virtualization optimization in software defined vehicular ad-hoc networks," in *Proc. of IEEE 84th Vehic. Tech. Conf. (VTC2016-Fall)*, Sept. 2016.

[7] I. Da Silva, S. E. El Ayoubi, O. M. Boldi, O. Bulakci, P. Spapis, M. Schellmann, H. ERC, J. F. Monserrat, T. Rosowski, *et al.*, "5G RAN architecture and functional design, METIS II White Paper.

[8] C. Niephaus, O. Aliu, M. Kretschmer, S. Hadzic, and G. Ghinea, "Wireless back-haul: a software defined network enabled wireless backhaul network architecture for future 5g networks, *Networks, IET*, vol. 4, no. 6, pp. 287-295, 2015.

[9] N. Zhang, N. Cheng, A. Gamage, K. Zheng, J. W. Mark, and X. Shen, "Cloud assisted HetNets toward 5G wireless networks," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 59-65, 2015.

[10] N. M. M. K. Chowdhury, "Network virtualization: State of the art and research challenges," *IEEE Communications Magazine*, vol. 47, no. 7, pp. 20-26, Jul. 2009.

[11] K. Tutschku, T. Zinner, A. Nakao, and P. Tran-Gia, "Network virtualization: Implementation steps towards the future internet," in *Proc. of the Workshop on Overlay and Network Virtualization at KiVS*, Kassel, Germany, Mar. 2009.

[12] C. Liang, and F. Richard Yu. "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 358-380, 2015.

[13] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," in *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pp. 1-9, IEEE, 2015.

[14] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path splitting and migration," *ACM SIGCOMM CCR*, vol. 38, no. 2, pp. 17-29, Apr. 2008.

[15] N. M. M. K. Chowdhury, M. R. Rahman, and R. Boutaba, "Virtual network embedding with coordinated node and link mapping," in *Proc. of IEEE INFOCOM*, 2009, pp. 783-791.

[16] A. Fischer, J. Botero, M. Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey." *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 1888-1906, 2013.

[17] D. Eppstein, "Finding the k shortest paths," *SIAM Journal on Computing*, vol. 28, pp. 652-673, Feb. 1999.

[18] M. Piro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004.

[19] E. Zegura, K. Calvert, and S. Bhattacharjee, "How to model an Internetwork," in *Proc. of IEEE INFOCOM*, 1996, pp. 594-602.