Optimal Distributed Resource Allocation in 5G Virtualized Networks

Hassan Halabian
Huawei Canada Research Center, 303 Terry Fox Dr., Kanata, ON, K2K 3J1, Canada
hassan.halabian@huawei.com

***Abstract*—The concepts of network function virtualization (NFV) and end-to-end (E2E) network slicing are two promising technologies empowering 5G networks for efficient, flexible and dynamic network deployment and service management. Optimal resource allocation is one of the challenging problems to address in such networks. In this paper, we propose a resource allocation model for 5G virtualized networks in a heterogeneous cloud infrastructure. In our model, each network slice has a resource demand vector for each of its building virtual network functions (VNFs). We then formulate the optimal resource allocation as a convex optimization problem maximizing the overall system utility as a function of the slice thicknesses with the constraints of the data centers' resource capacities. The slice thickness variables together with the demand vectors determine the amount of resources allocated to each slice. We further propose a distributed solution for the resource allocation problem based on auction/game theory by forming a resource auction between the slices and the data centers (DCs). It is shown that the resource allocation game has a unique Nash equilibrium and its solution is the same as the solution of the centralized system optimization problem, i.e., in equilibrium the slice thicknesses maximize the overall system utility. Numerical analysis are provided to show the validity of the results, evaluate the convergence of the distributed solution and also comparing the performance of the optimal scheme with heuristic ones.**

***Index Terms*—5G Network Function Virtualization, Network Slicing, Resource Allocation, Algorithmic Games**

I. Introduction

Network Function Virtualization (NFV) and Software-Defined Networking (SDN) are two promissing techniques used in 5G network architecture evolution to provide significant capital and operational expenditure saving by immigrating the network functions and services to cloud infrastructures [1]–[4]. NFV provides software and hardware decoupling by virtualizing the service components and network functions and running them on top of a virtualization system, i.e., virtual machines or containers [5]. On the other hand, SDN provides centralized control plane for control and management of network services and network functions. These two techniques together with the concept of E2E network slicing [6]–[9] enable mobile network providers to create virtualized E2E networks over cloud systems. Depending on the functional, operational and performance requirements, there have been defined a number of 5G network slices in accordance with the concept of networks as a service (NaaS), including but not limited to enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (uRLLC) and massive machine-type communications (mMTC) [10].

978-3-903176-15-7

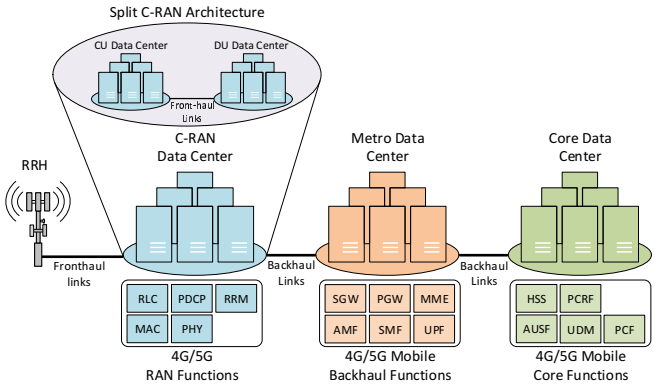

Fig. 1: 5G network fuction virtualization - C-RAN and mobile backhaul architecture

A virtualized network slice consists of a number of VNFs distributed geographically in numerous DCs. Each VNF provides certain services in its slice and all the VNFs of a slice collectively provide wireless network access to the UEs attached to that slice. Fig. 1 shows an illustration of network function virtulization architecture for 5G networks which provides RAN (Radio Access Network) and mobile backhaul/core function virtualization in data centers. As shown in Fig. 1, in 5G C-RAN (Cloud-Radio Access Network) architecture, communication signals are collected from the cell towers by the Remote Radio Heads (RRH) and after RF (Radio Frequency) processing they are sent to the Base Band Units (BBUs) for digital processing. BBU may itself split into Central Unit (CU) and Distributed Unit (DU) [11]. DU runs latency sensitive RAN functions while CU is supposed to run latency tolerant functions. In C-RAN architecture, some or all BBU RAN functions may be virtualized [12]. Packet level processing is done in SGW (Serving Gateway) and PGW (Packet Gateway) and mobility services are provided by MME (Mobility Management Entity). Subscriber related information processing, e.g., authentication, location, etc., is done by HSS (Home Subscriber Server) [13]. In 4G technology, these functions are implemented in dedicated hardware while in a virtulaized architecture they are placed as virtual machines/containers in DCs promoting the concept of Mobile Carrier Cloud [4]. For 5G networks, 3GPP defines a service-based network architecture in which mobile

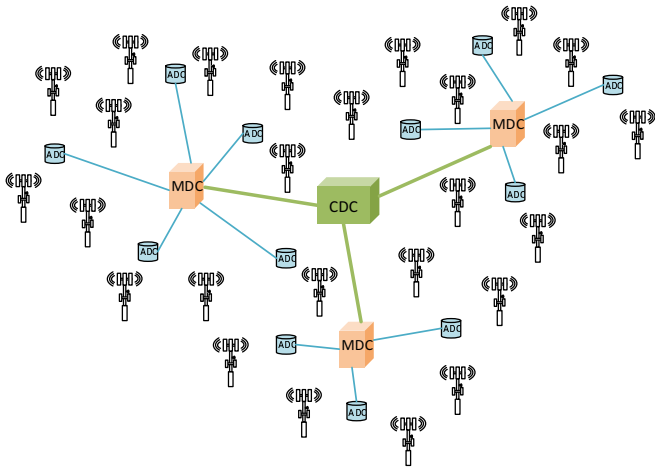

Fig. 2: A topology of ADC, MDC and CDC distribution

back-haul/core services are provided by virtualized network functions including (but not limited to) Access and Mobility Management Function (AMF), Session Management Function (SMF), Authentication Server Function (AUSF), User Plane Function (UPF), Unified Data Management (UDM) and Policy Control Function (PCF) [14].

Due to interdependence of VNFs in each slice, there are placement constraints for placing the VNFs in the serving DCs. For instance, due to front-haul latency and jitter constraints, there is distance limitation between the radio processing functions and the BBU functions. Furthermore, certain slices, e.g., uRLLC, require to satisfy certain QoS requirements and therefore there exist some placement constraints for the VNFs of those slices. There might also exist some placement restrictions due to administration, logistics and management concerns. In the model presented in this paper, it is assumed that the VNFs are placed in predetermined DCs called Access Data Center (ADC), Metro Data Center (MDC) and Core Data Center (CDC). ADC hosts VNFs processing L1/L2 access functions, e.g., RLC, PDCP, RRM, MAC and PHY. ADCs are preferably located physically close to the cell sites and RRHs due to latency considerations. MDC hosts VNFs processing traffic forwarding, classifications, admission and mobility management, etc. Examples of functions in MDC are 4G functions MME, SGW, PGW and 5G functions AMF, SMF and UPF. CDC may locate functions dealing with subscriber related information and policy enforcement and charging functions. HSS is an example of a 4G function in CDC. In 5G systems, CDC may locate functions such as UDM and AUSF. Fig. 2 shows a topology illustration of the distribution of ADC, MDC and CDC data centers in a 5G network architecture. Similar modeling had been captured for 5G packet core in the literature, e.g., in [8].

The VNFs of a single slice have heterogeneous resource requirements, i.e., CPU, memory, bandwidth and storage. For example, BBU functions are CPU intensive as they execute heavy processing DSP functions while PGW is a bandwidth intensive function as it passes the entire slice traffic. The resource requirements of slices of the same type are also different since they are serving different number of UEs. For instance a provider might run multiple IoT (Internet of Things) slices each one dedicated for a specific application [15], [16]. These slices might have different resource demands depending on the number of attached UEs to them and also the type of IoT application. Furthermore, the serving DCs have heterogeneous resource capacities for each of their resources. For such a system, with heterogeneous resource capacities and heterogeneous slice requirements, optimal resource allocation to the network slices is a challenging problem. Each network slice might have a different utility/revenue function not willing to share with DCs. Moreover, the DCs might not be under the same management. For such a model, a distributed resource allocation scheme is more preferable for both the slice providers and also the DCs.

Our contributions in this paper are summarized as follows: We first propose a resource allocation model for 5G virtualized networks in a heterogeneous cloud infrastructure with E2E network slices having diverse requirements and resource demands. The heterogeneity of the slice requirements is reflected in our model by considering different resource demand vectors for each function of each slice. The demand vectors for each slice specify the amount of resources required for each function to complete a network task in one unit of time, e.g., to serve one wireless user equipment. Hence, the resource volume of each slice can be specified by a scalar multiplier of its demand vectors and is called *slice thickness* in this paper. The resource allocation optimization is maximizing the total network utility as a function of the slice thicknesses with the constraints of the DCs' resource capacities. The utility functions are assumed to be strictly concave and thus the resource allocation is a convex optimization problem. Specific choices of utility functions may provide desired fairness properties, e.g., max-min, α-proportional, etc. We further present a distributed solution for solving the resource allocation problem by forming a resource allocation auction between the slices and the DCs. It is proven that the resource allocation game has a Nash equilibrium and also the Nash solution is the same as the solution of the centralized system optimization problem, i.e., in equilibrium the slice thicknesses are also maximizing the overall system utility function. Numerical analysis are provided to support the validity of the results.

The rest of the paper is organized as follows. Section II, presents the related research work in this domain of research. In Section III, the proposed 5G resource allocation model is introduced. Section IV formulates the global centralized system utility optimization problem. In Section V, we present the distributed game-based solution and show the existence of Nash equilibrium. We further show that the Nash solution coincides with the optimal system utility resource allocation. Section VI, presents the simulation and numerical results. Section VII provides the conclusions of the paper.

II. Related Work

The concept of network slicing provides flexible and dynamic provisioning of network services to vertical industries including but not limited to manufacturing, health care, media and entertainment, automotive, public safety, financial services, etc. [17]. The research in this area resulted in forming many joint projects over open source platforms such as OPNFV [18] and OpenMANO [19] for management and orchestration of wireless network functions [1], [2].

Many research activities in this area are mainly focused on radio resource virtualization [6], [7], management and orchestration of network functions [8], [9] without considering the heterogeneous demand, QoS and performance requirements of slices. The work in [20] focuses on Evolved Packet Core (EPC) virtualization and addresses the optimal placement of SGW and PGW functions in cloud carrier networks without considering the end-to-end network slice requirements.

While most of the work in VNF orchestration is dedicated to per DC orchestration of the VNFs, service-centric slice orchestration (with diverse E2E requirements) is studied far less in the literature [21]. In [11], [21]–[23], there have been proposed algorithms for optimal VNF resource allocation problem. In [23], the authors formulate a mixed-integer linear programming (MILP) for joint function chaining and resource allocation problem and to solve this problem they propose heuristic alternatives. Similarly [22] formulates the function chaining problem as a binary NP-hard programming problem and to solve it the authors propose heuristic approaches. In [21], complex network theory is used to obtain topological information of slices and infrastructure network and ranking the nodes for mapping VNFs to the nodes. In none of these papers, the model is comprehensive in a sense to consider the DC models and the available resources in DCs (computation, memory and bandwidth) in the problem formulation. Moreover, the objective in all of them misses the slice provider utilities, fairness and also heterogeneous resource demands of the network slices. In [24], the authors support the idea of high-level system orchestration for dynamic management of VNFs and propose an architectural system model without proposing a technical function placement and resource allocation solution. The work in [11] formulates a MILP to derive the optimal number of VNFs to meet the performance requirements of a network slice. The authors further form a coalition game between DCs to host the slice VNFs. In contrast to the aforementioned references, we consider the resource allocation among a number of competing slices with diverse resource demands on a set of heterogeneous DCs. Moreover, we formulate the resource allocation with the objective of maximizing the overall system utility. Our distributed scheme forms an auction game between the slices and the DCs (in contrast to [11] where the game is between the DCs) to solve the system optimization problem.

III. System Model

Consider a virtualized 5G system consisting of a set of N network slices. Each slice n is composed of a number of VNFs denoted by $\mathcal{F}^n = \{f^n_{\mathbf{1}}, f^n_{\mathbf{2}}, \cdots, f^n_M\}$ where M is the number of VNFs for each slice. If different network slices have different number of VNFs, we let M be the maximum. Fig. 3 shows an illustration of the system model for three slices.

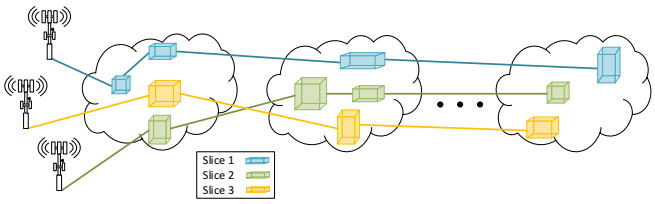

Fig. 3: System model for function placement and resource allocation

In our model, there are K DCs over which the VNFs are being distributed. The available resources in each DC k are denoted by vector $\mathbf{R}_k = (R_{k,\mathbf{1}}, R_{k,\mathbf{2}}, \cdots, R_{k,L})$ where $R_{k,\ell}$ represents the amount of available resource ℓ on DC k. The available resources in each DC are for example CPU, memory, bandwidth and storage. Each VNF f^n_m of slice n is placed in one of the DCs. The placement of each VNF in DCs is predetermined by the operator. However, the amount of resources allocated to each VNF in each DC is unknown. The goal of the resource allocation is to determine the amount of resources allocated to each of the VNFs in each DC.

We assume that each network slice n is associated with a set of demand vectors for each of its VNFs denoted by $\mathbf{d}^k_{f^n_m} = (d^k_{f^n_m,\mathbf{1}}, d^k_{f^n_m,\mathbf{2}}, \cdots, d^k_{f^n_m,L})$ for each DC k and each VNF m of slice n, denoted by f^n_m. If VNF m for slice n (i.e., f^n_m) is not defined or is not going to be placed on DC k, we set $\mathbf{d}^k_{f^n_m}$ to zero vector. The set of demand vectors of each network slice is denoted by $\mathcal{D}_n = \{\mathbf{d}^k_{f^n_m} | m = 1, 2, \cdots M, k = 1, 2, \cdots, K\}$ which reflects the amount of resources for each VNF in each DC to complete a network task in one unit of time, e.g., to serve one wireless user equipment. The demand vectors are heterogeneous across the network slices and also inside each network slice. We now define the slice thickness variable v_n ($v_n > 0$) for each slice n which denotes the number of network tasks (wireless user service) that can be executed in one unit of time. We can interpret $\mathcal{D}_n$ as the slice thickness for a single task and therefore $v_n\mathcal{D}_n$ represents the amount of resources allocated to slice n to support v_n task(s) in one unit of time. In other words, v_n specifies how the network slice expands or shrinks with respect to its demand vectors $\mathcal{D}_n$. By $\mathbf{v} = (v_{\mathbf{1}}, v_{\mathbf{2}}, \cdots, v_N)$ we denote the vector of the slice thicknesses. To have a visual view of the problem, consider Fig. 3 again. In this figure, each cube represents the demand vector for each VNF such that each side magnifies the demand element for each resource type. The cubes of the same color are the building VNFs of a network slice. A slice thickness variable v_n determines how the cubes of the same slice expand or shrink with v_n as the expansion coefficient.

We assume that slice operators have separate utility functions as a function of the amount of resources allocated to the slice's functions. Since the allocations of each slice scale

with the slice thickness v_n, the slice n utility function can be denoted by $U_n(v_n)$. It is assumed that the slice utility function is increasing, strictly concave and continuously differentiable function of v_n. These assumptions on the utility functions are realistic assumptions as each slice utility/revenue will be increasing with respect to its allocated resources but the slope of utility growth decreases by increasing the allocations, e.g., due to limited number of UEs. The problem to address here is to find the slice thickness variables v_n such that the total system utility is optimized.

IV. Centralized Resource Allocation Optimization

Based on the assumptions and the presented system model, we can formulate the virtualized 5G resource allocation into the following optimization problem.

Centralized System Optimization:

$$\begin{aligned} \underset{\mathbf{v}}{\texttt{Maximize:}} \quad & \sum_{n=1}^{N} U_n(v_n) && (1)\\ \texttt{Subject to:} \quad & \sum_{n=1}^{N}\sum_{m=1}^{M} v_n d^k_{f^n_m,\ell} \leq R_{\ell,k} \quad \forall \ell, \forall k \\ & v_n \geq 0 \quad \forall n \end{aligned}$$

The objective of Problem (1) is to maximize the overall system utility defined as the sum of the slices' utility functions. The constraints of this problem ensure that the slice thickness allocations will not violate the capacity limits of each resource in each DC. The centralized system optimization problem is a convex optimization problem in terms of the slice thickness variables v_n. This is because the objective function is concave and the constraints are linear inequalities representing a compact feasible region. Therefore, the centralized optimization problem has a unique optimum solution [25].

By choosing a proper utility function $U_n(\cdot)$, we can achieve a trade-off between efficiency and fairness which depends on the specific choice of $U_n(\cdot)$. To capture the trade-off between efficiency and fairness, one may choose $U_n(\cdot)$ from the class of α-fair utility functions [26], [27]. Specifically, by choosing $U_n(\cdot)$ such that $U'_n(x) = x^{-\alpha}$, for some fixed parameter α, the optimal solution of the centralized system optimization satisfies α-proportional fairness in terms of slice thicknesses. Thus, the α-proportional utility function is defined by $U_n(v_n) = \frac{v_n^{1-\alpha}}{1-\alpha}$ for $\alpha > 0$ and $\alpha \neq 1$. For $\alpha = 1$, we have $U_n(v_n) = \log(v_n)$ which is equal to the limiting value of $\frac{v_n^{1-\alpha}}{1-\alpha}$ when $\alpha \to 1$. The α-proportional utility function with $\alpha = 1$ provides "proportional fair" allocation and when $\alpha \to \infty$, it provides "max-min" fairness.

The centralized system problem can be solved by well-known convex optimization methods, e.g., subgradient projection and interior-point methods by a central optimizer [25]. The main issues of centralized solutions are lack of scalability and single-point-of-failure problem. With the growth of the number of network slices and their VNFs and dynamic network changes, scalability becomes an important challenge for centralized solutions. Moreover, any failure in the central optimizer may result in the entire resource allocation scheme to fail. Another drawback of centralized approaches is that the slice providers need to disclose their (possibly private) utility functions with DCs. Finally, centralized solutions fail to provide a global optimal resource allocation if the DCs are not under the same management and do not want to disclose their resource capacities to a third party optimizer.

V. Distributed Resource Allocation - Auction Game Approach

Due to the problems with centralized approaches, we propose the following distributed scheme for 5G resource allocation problem. The distributed scheme is based on application of auction theory by forming an auction game between the slices and the DCs. In this scheme, the network slices bid for each of the resources of the DCs they are placing a function on. Based on the bids submitted by the network slices, the price for each resource on each DC is determined and is announced to the network slices together with their calculated thickness values. Each slice thickness will be equal to the minimum of the slice thicknesses received from all DCs. On the other hand, each slice maximizes its payoff based on the prices received from the DCs and updates its bid for the next round of the game. It is shown that Nash equilibrium exists for such an auction, i.e., there exists an equilibrium slice thickness vector and an equilibrium resource price for each of the DCs' resources such that no network slice is willing to change its bid and its allocation. Furthermore, it is shown that Nash solution for the game problem is the same as the solution of the centralized system optimization Problem (1), i.e., the Nash equilibria of the game approach will achieve the full efficiency of the system optimization. The benefits of the proposed distributed scheme are the following:

- Convergence to the system optimal solution.
- No optimization third party involved and no information sharing between the slice providers and the DCs.
- DCs do not necessarily need to be under the same management. This provides flexibility for the slice provider to choose proper DCs for placing its functions, i.e., flexibility for different business models.

A. Game setup

The resource allocation game is setup in the following items:

- Each Slice n offers an amount of $w^k_{f^n_m,\ell}$ for resource ℓ of DC k which is locating the VNF f^n_m. The bidding is done for all ℓ, k and m. We define $\mathbf{w} = (w^k_{f^n_m,\ell}, \forall n, \forall m, \forall k, \forall \ell)$ as the offer matrix.
- Each DC k calculates the price of each of its resources by using the following equation.

$$p_{\ell,k} = \frac{\sum_{n,m} w^k_{f^n_m,\ell}}{R_{\ell,k}} \qquad (2)$$

We define $\mathbf{p} = (p_{\ell,k}, \forall \ell, \forall k)$ as the resource price matrix.

- Each DC k calculates a slice thickness for each slice n for each VNF f_m^n and resource ℓ as follows.

$$v^k_{f^n_m,\ell} = \frac{w^k_{f^n_m,\ell}}{p_{\ell,k} d^k_{f^n_m,\ell}} \tag{3}$$

We call these thicknesses as the *deficient thicknesses* since they are calculated just based on local and insufficient information for each resource of each DC.

- The resource price on each DC as well the deficient thicknesses are announced to the slices. Each slice calculates the offered final thickness by

$$v_n = \min_{m,\ell,k} \{v^k_{f^n_m,\ell}\}. \tag{4}$$

- Each slice further uses the resource prices to update its thickness by maximizing its overall payoff function $G_n(v_n; \mathbf{p})$ defined as

$$G_n(v_n; \mathbf{p}) = U_n(v_n) - v_n \sum_{\ell,m,k} d^k_{f^n_m,\ell} p_{\ell,k}. \tag{5}$$

Slice n Payoff Optimization:

$$\begin{aligned} &\texttt{Maximize:} && G_n(v_n; \mathbf{p}) \\ &\texttt{Subject to:} && v_n \geq 0 \end{aligned} \tag{6}$$

- Each slice then updates its offer for each resource to each DC for each of its functions by

$$w^k_{f^n_m,\ell} = v^\star_n d^k_{f^n_m,\ell} p_{\ell,k}, \tag{7}$$

where $v^\star_n$ is the solution of the slice payoff optimization problem in (6).

The game is considered to be converged if the distance of the thickness allocation from the DCs derived from (4) and the thickness derived from the slice payoff optimization problem in (6) is less than ϵ, i.e., *for all* n, $|v_n - v^\star_n| \leq \epsilon$ where ϵ is a given parameter of the algorithm and establishes a trade-off between the solution accuracy and the convergence speed. We now formally define the Nash equilibrium for the aforementioned game.

Definition 1. *The game is in Nash equilibrium if there exists a pair of offer matrix and resource price matrix* $(\mathbf{w}, \mathbf{p})$ *such that the slice payoff defined in (6) is maximized and the equilibrium resource price is determined according to (2), i.e.,*

$$G_n(v_n; \mathbf{p}) \geq G_n(\bar{v}_n; \mathbf{p}) \quad \text{for any} \quad \bar{v}_n \geq 0, \quad \forall n \tag{8}$$

$$p_{\ell,k} = \frac{\sum_{n,m} w^k_{f^n_m,\ell}}{R_{\ell,k}} \qquad \forall \ell, \ \forall k. \tag{9}$$

In the following theorem, we show that the Nash equilibrium does exist for the described game and the resulting Nash thickness vector is equal to the solution of the centralized system optimization in (1).

Theorem 1. *Assume that the slice utility functions are strictly concave, increasing and continuously differentiable. Nash equilibrium exists for the resource allocation game described above, i.e., there exists a pair* $(\mathbf{w}, \mathbf{p})$ *such that (8) and (9) are satisfied. Furthermore, the Nash pair* $(\mathbf{w}, \mathbf{p})$ *will result in a unique thickness vector* $\mathbf{v}$ *derived from (4) such that it also solves the centralized system optimization in (1).*

Proof. The proof follows by taking the Lagrangian of the optimization problem and showing that the Nash conditions (8) and (9) are the same as the optimality conditions of the centralized system optimization problem as used in [27]. Since the centralized system optimization problem is strictly feasible (at least $\mathbf{v} = 0$ is in the feasible region) then Slater condition guarantees that the strong duality for this problem holds [25]. Also since the objective function is strictly concave, increasing and continuously differentiable and the feasible region is compact, the solution is unique [25]. The Lagrangian form for this problem is given by

$$L(\mathbf{v}; \lambda) = \sum_{n=1}^{N} U_n(v_n) - \sum_{\ell,k} \lambda_{\ell,k} \left(\sum_{n=1}^{N} \sum_{m=1}^{M} v_n d^k_{f^n_m,\ell} - R_{\ell,k} \right) \tag{10}$$

where λ is the matrix of Lagrangian variables. Assuming that $\mathbf{v}^*$ is the optimal vector of slice thicknesses for Problem (1), KKT (Karush-Kuhn-Tucker) conditions ensure that there exist Lagrange multipliers $\lambda_{\ell,k}$ such that the following conditions (primal and dual feasibility, complementary slackness and vanishing of the gradient of the Lagrangian) are hold [25].

$$\sum_{n=1}^{N} \sum_{m=1}^{M} v^*_n d^k_{f^n_m,\ell} \leq R_{\ell,k}, \quad \forall \ell, \ \forall k \tag{11}$$

$$v^*_n \geq 0, \quad \forall n \tag{12}$$

$$\lambda_{\ell,k} \geq 0 \quad \forall \ell, \ \forall k \tag{13}$$

$$\lambda_{\ell,k} \left(\sum_{n=1}^{N} \sum_{m=1}^{M} v^*_n d^k_{f^n_m,\ell} - R_{\ell,k} \right) = 0 \quad \forall \ell, \ \forall k \tag{14}$$

$$v^*_n > 0 \Rightarrow U'_n(v^*_n) = \sum_{\ell,m,k} \lambda_{\ell,k} d^k_{f^n_m,\ell} \qquad \forall n \tag{15}$$

$$v^*_n = 0 \Rightarrow U'_n(v^*_n) \leq \sum_{\ell,m,k} \lambda_{\ell,k} d^k_{f^n_m,\ell} \qquad \forall n \tag{16}$$

The Nash equilibrium is at a point where given the resource prices, the payoff of all the slices are maximized, i.e., for all n,

$$G'_n(v_n; \mathbf{p}) = U'_n(v_n) - \sum_{\ell,m,k} p_{\ell,k} d^k_{f^n_m,\ell} \leq 0, \quad v_n \geq 0 \tag{17}$$

$$v_n(G'_n(v_n; \mathbf{p})) = 0 \Rightarrow v_n \left(U'_n(v_n) - \sum_{\ell,m,k} p_{\ell,k} d^k_{f^n_m,\ell} \right) = 0 \tag{18}$$

On the other hand, the equilibrium resource prices satisfy either of the following equations:

$$\begin{aligned} &\sum_{n=1}^{N} \sum_{m=1}^{M} v_n d^k_{f^n_m,\ell} = R_{\ell,k} \\ \text{or} \quad &\sum_{n=1}^{N} \sum_{m=1}^{M} v_n d^k_{f^n_m,\ell} \leq R_{\ell,k} \text{ and } p_{\ell,k} = 0 \end{aligned} \tag{19}$$

The first equation says that with the current price and thicknesses the capacity $R_{\ell,k}$ is totally used up. The second one

says that if the capacity is not used up (i.e., there is no competitive demand for it in the auction) its price tends to zero. Note that in equilibrium, since the thickness v_n is determined from (4), the overall demand for some of resources on some DCs might not be fully demanded, i.e., $\sum_{n=1}^{N}\sum_{m=1}^{M} v_n^* d^k_{f_m^n,\ell} \leq R_{\ell,k}$ and thus for those resources, the price $p_{\ell,k}$ tends to zero. Conditions (19) can be summarized as

$$p_{\ell,k}\left(\sum_{n=1}^{N}\sum_{m=1}^{M} v_n d^k_{f_m^n,\ell} - R_{\ell,k}\right) = 0,$$
$$p_{\ell,k} \geq 0, \quad \sum_{n=1}^{N}\sum_{m=1}^{M} v_n d^k_{f_m^n,\ell} \leq R_{\ell,k}. \quad (20)$$

We observe that Conditions (11) – (16) are similar to Conditions (17) – (20). Since the primal and dual centralized system optimization have unique solutions $\mathbf{v}^*$ and λ, by letting $p_{\ell,k} = \lambda_{\ell,k}$, we observe that the equilibrium price matrix $\mathbf{p}$ does exist and is unique. We now show that the Nash equilibrium point in the resource allocation game is the same as the optimal point in the centralized system optimization problem. From (6) at the equilibrium, we know that

$$U_n(v_n) - v_n \sum_{\ell,m,k} d^k_{f_m^n,\ell} p_{\ell,k} \geq U_n(v_n^*) - v_n^* \sum_{\ell,m,k} d^k_{f_m^n,\ell} p_{\ell,k} \quad (21)$$

By summing over all slices,

$$\sum_n U_n(v_n) \geq \sum_n U_n(v_n^*) + \sum_{n,\ell,m,k} v_n d^k_{f_m^n,\ell} p_{\ell,k} - \sum_{n,\ell,m,k} v_n^* d^k_{f_m^n,\ell} p_{\ell,k} \quad (22)$$

From (20), we have

$$\sum_{n,\ell,m,k} v_n d^k_{f_m^n,\ell} p_{\ell,k} = \sum_{\ell,k} p_{\ell,k} R_{\ell,k} \quad (23)$$

From the constraints of Problem (1), we also have

$$\sum_{n,\ell,m,k} v_n^* d^k_{f_m^n,\ell} p_{\ell,k} \leq \sum_{\ell,k} p_{\ell,k} R_{\ell,k} \quad (24)$$

Using (22) – (24), we observe that

$$\sum_n U_n(v_n) \geq \sum_n U_n(v_n^*) \quad (25)$$

Since, $\mathbf{v}^*$ is the global optimal point for the centralized system Problem (1), we must have $\sum_n U_n(v_n) = \sum_n U_n(v_n^*)$ and since the solution is unique, we have $\mathbf{v} = \mathbf{v}^*$. □

In summary, Theorem 1 proves the following statements:

- A unique Nash equilibrium exists for the proposed auction game, i.e., there exist an equilibrium slice thickness vector and an equilibrium resource price for each of the DCs' resources such that no network slice is willing to change its bid and its allocation.
- Nash solution for the game problem is equal to the solution of the centralized system optimization problem in (1), i.e., the gap between the game result in Nash equilibrium and the solution of Problem (1) is zero. Hence, the game approach achieves the full efficiency of the system optimization.

TABLE I: Data centers resource settings

DC	CPU (cores)	RAM(GB)	BW(Gbps)	Storage(TB)
1	5000	10000	1000	2000
2	5000	5000	2000	5000
3	5000	5000	2000	10000

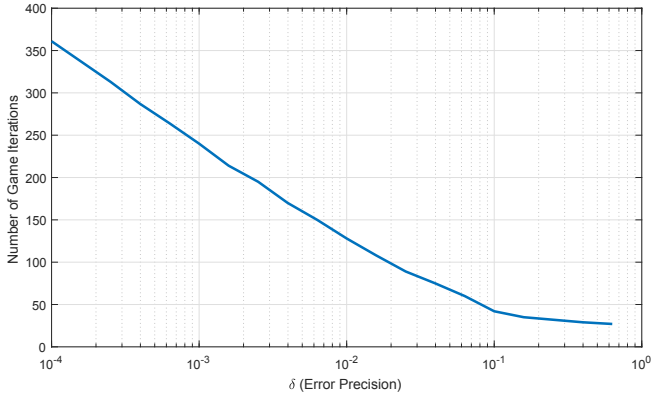

Fig. 4: Convergence of the distributed scheme

VI. Simulation Results

In this section, we present our simulation/numerical results in which we first confirm the convergence of the distributed scheme to the optimal thickness vector via numerical analysis over sample system setups. We then compare the performance of different α-proportional fairness utility functions in terms of resource allocation efficiency. Finally, we compare the optimal solution with the solution of two heuristic schemes in terms of system utility and resource utilization. Recall that the α-proportional utility function is defined by $U_n(v_n) = \frac{v_n^{1-\alpha}}{1-\alpha}$ for $\alpha > 0, \alpha \neq 1$ and $U_n(v_n) = \log(v_n)$ for $\alpha = 1$.

We consider a system consisting of three DCs and 100 network slices. Each network slice is composed of 5 VNFs. VNF 1 is placed at DC 1, VNFs 2 and 3 are placed at DC 2 and VNFs 4 and 5 are placed at DC 3 for all slices. Each DC contains 4 types of resources, CPU, memory, network bandwidth and storage. Table I shows the amount of available resources in each DC. The elements of demand vectors of all network slices, for all the functions are randomly selected as follows: for CPU, from the interval [1 10] cores; for RAM, from the interval [1 10] GB; for storage, from the interval [1 10] TB and for network bandwidth, from the interval [0.25 2.5] Gbps.

A. Convergence of the Distributed Scheme

To show the convergence of the distributed algorithm, we consider a system in which each slice has an α-proportional fair utility function where α is chosen randomly for each slice from the interval 1 to 10. Recall that we introduced ϵ in Section V-A as the precision parameter for the convergence of

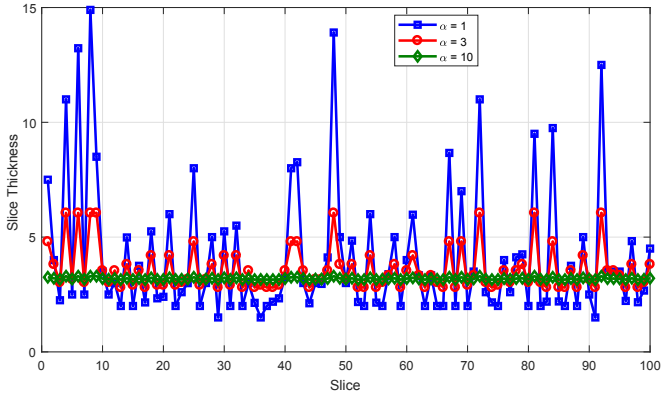

Fig. 5: Slice thickness for α-proportional fair utility functions ($\alpha = 1, 3, 10$)

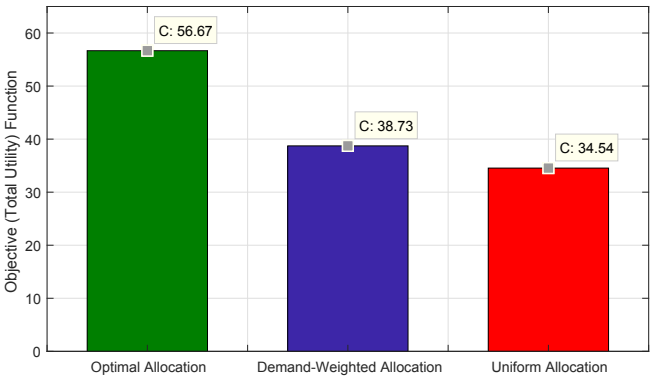

Fig. 7: Objective values - optimal and heuristic schemes

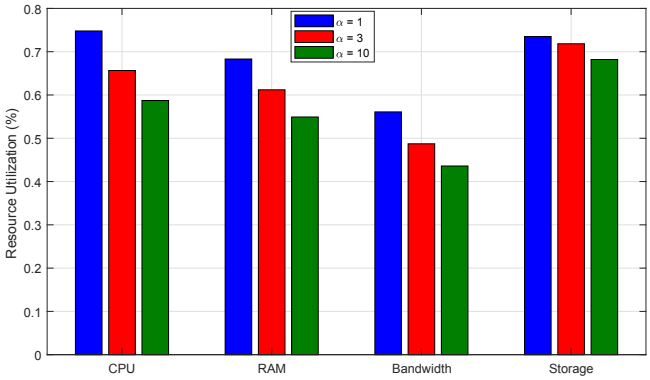

Fig. 6: Resource utilization for α-proportional fair utility functions ($\alpha = 1, 3, 10$)

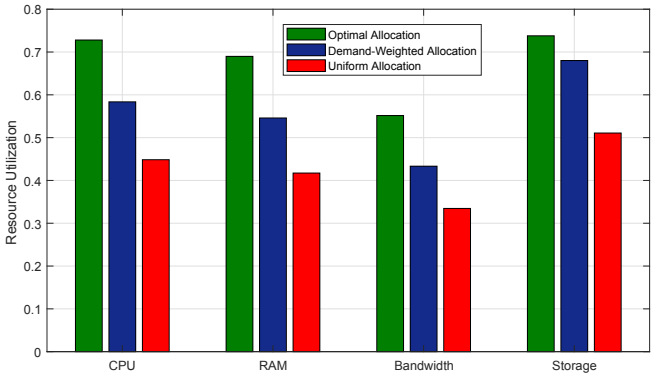

Fig. 8: Resource utilization comparison for the optimal and the heuristic schemes

the resource allocation game. In Fig. 4, we have measured the number of game iterations required for the system to converge with precision ϵ for sample values of ϵ ranging from 0.5 to 10^{-4}. It is observed that for precision $\epsilon = 0.1$ which is a reasonable precision value, less than 50 iterations is enough for the system to converge. Moreover, it is observed that the required number of iterations is decreasing linearly as the precision error grows logarithmically meaning that we can achieve sophisticated precision errors efficiently by linearly increasing the number of iterations.

B. Resource Efficiency for Different α-proportional Utilities

In this section, we compare the allocation and resource utilization under 3 different α-proportional fairness utility functions. We assume all the the network slices have the same α-proportional fairness utility with $\alpha = 1, 3, 10$. Fig. 5 shows the slice thickness values for each slice for different values of α. With $\alpha = 1$, allocations are proportionally fair, i.e., the system tries to maintain a balance between fairness and resource utilization. By increasing the α parameter, the fairness behavior of the system tends to max-min fair allocation where resource utilization is ignored and the objective is only maximizing the minimum allocation among the network slices. Fig. 6 shows the resource utilization for each α. Note that the resource utilization is measured for each resource per DC as the ratio of the total allocation of each resource and the available amount of that resource in the DC and then it is averaged over the three DCs. It is observed that with $\alpha = 10$, the thickness allocations are almost equal for every slice while the resource utilization efficiency is the least. However, with $\alpha = 1$ there are fluctuations in the thickness allocations and the resource utilization is the maximum for all types of resources.

C. Comparison with Heuristic Sub-optimal Schemes

In this section, we compare the performance of the optimal distributed scheme with two heuristic schemes. The first scheme allocates the resources of each DC uniformly among its VNFs. In this scheme, not all resources allocated to a network slice are useful for it. The effective utilized resources for each slice depends on its allocations as well as its demand vectors. We call this allocation as the uniform allocation. The second scheme allocates the available resources to the VNFs based on the demand vectors of the VNFs, i.e., the allocations are weighted based on the elements of the demand vectors of the VNFs for each resource type and for each DC. The resultant allocation for this scheme is such that all network slices will get the same slice thickness. This

allocation is called the demand-weighted scheme. Note that in both heuristic schemes, the maximal information needed for resource allocation is the demand vectors and they operate independently of the utility functions. This assumption is due to the fact that slice providers are reluctant to share their private information with DCs.

For this scenario, we again assume that each network slice has an α-proportional utility function with a randomly selected α from 1 to 10 for each network slice. Fig. 7 compares the objective value of the resource optimization Problem (1) for the optimal distributed scheme and the two heuristics. Fig. 8 shows the resource utilization comparison among the optimal and the heuristic approaches. It is observed that the optimal allocation results in the maximum overall system utility and also outperforms the heuristic ones in terms of resource utilization.

VII. Conclusions

We have introduced a model of resource allocation for 5G networks incorporating the notions of network function virtualization and end-to-end network slicing. We formulated the optimal resource allocation as a convex problem with the objective to maximize the overall system utility function as a function of the slice resource allocations indicated by slice thickness variables. We introduced a distributed auction-based approach to solve the system optimization problem and showed theoretically that the auction game has a unique Nash equilibrium and also it converges to the global optimal system solution. Simulation results were provided to evaluate the performance of the distributed scheme in terms of convergence and resource utilization for different utility functions. We also compared its performance with two heuristic approaches.

References

[1] ETSI and GSNFV, "network functions virtualization (nfv); architectural framework," ETSI, Oct. 2013.

[2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236 – 262, First quarter 2016.

[3] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, Third quarter 2017.

[4] T. Taleb, "Towards carrier cloud: Potential, challenges & solutions," *IEEE Wireless Comm. Mag.*, vol. 21, no. 3, pp. 80 – 91, Jun. 2014.

[5] A. Laghrissi, T. Taleb, and M. Bagaa, "Conformal mapping for optimal network slice planning based on canonical domains," *IEEE Journal on Selected Areas in Communications (Early Access)*, 2018. [Online]. Available: 10.1109/JSAC.2018.2815436

[6] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network store: Exploring slicing in future 5g networks," in *International Workshop on Mobility in the Evolving Internet Architectur*, Paris, France, Sep. 2015.

[7] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Comm. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[8] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5g," *IEEE Comm. Mag.*, vol. 54, no. 5, pp. 84–91, May 2016.

[9] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5g network slice broker," *IEEE Comm. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[10] M Series, "IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond," International Telecommunication Union (ITU), Sep. 2015.

[11] M. Bagaa, T. Taleb, and A. Laghrissi, "Coalitional game for the creation of efficient virtual core network slices in 5g mobile systems," *IEEE Journal on Selected Areas in Communications (Early Access)*, 2018. [Online]. Available: 10.1109/JSAC.2018.2815398

[12] "C-ran: The road towards green ran-v2.5," China Mobile White Paper, Oct. 2011.

[13] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "Ease: Epc as a service to ease mobile core network," *IEEE Network Mag.*, vol. 29, no. 2, pp. 78 – 88, Mar. 2015.

[14] 3GPP, "3gpp ts 23.501: Technical specification group services and systems aspects; system architecture for the 5g system; stage 2," Rel. 15, V15.2.0, Jun. 2018.

[15] A. Ghasempour, "Optimum number of aggregators based on power consumption, cost, and network lifetime in advanced metering infrastructure architecture for smart grid internet of things," in *13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, Jan. 2016.

[16] A. Ghasempour and T. K. Moon, "Optimizing the number of collectors in machine-to-machine advanced metering infrastructure architecture for internet of things-based smart grid," in *IEEE Green Technologies Conference (GreenTech)*, Kansas City, MO, USA, Apr. 2016.

[17] Ericsson, Huawei, and Nokia, "5g network slicing for vertical industries," Global mobile Suppliers Association, Sep. 2017.

[18] Linux Foundation, "Opnfv." [Online]. Available: https://www.opnfv.org

[19] ETSI, "Open Source Mano." [Online]. Available: http://osm.etsi.org

[20] M. Bagaa, T. Taleb, and A. Ksentini, "Service-aware network function placement for efficient traffic handling in carrier cloud," in *IEEE WCNC'14*, Istanbul, Turkey, May 2014.

[21] W. Guan, X. Wen, L. Wang, Z. Lu, and Y. Shen, "A service-oriented deployment policy of end-to-end network slicing based on complex network theory," *IEEE Accesse*, vol. 6, pp. 19 691 – 19 701, Apr. 2018. [Online]. Available: 10.1109/ACCESS.2018.2822398

[22] J. Liu, Y. Li, Y. Zhang, L. Su, and D. Jin, "Improve service chaining performance with optimized middlebox placement," *IEEE Trans. on Services Computing*, vol. 10, no. 4, pp. 560 – 573, July - Aug 2017.

[23] L. Wang, Z. Lu, X. Wen, R. Knopp, and R. Gupta, "Joint optimization of service function chaining and resource allocation in network function virtualization," *IEEE Access*, vol. 4, pp. 8084 – 8094, Nov. 2016.

[24] S. Clayman, E. Maini, A. Galis, A. Manzalini, and N. Mazzocca, "The dynamic placement of virtual network functions," in *Network Operations and Management Symposium (NOMS)*, Krakow, Poland, May 2014.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[26] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237 – 252, Mar. 1998.

[27] F. P. Kelly, "Charging and rate control for elastic traffic," *Euro. Trans. Telecommun.*, vol. 8, pp. 33 – 37, Jan./Feb. 1997.