

Resource Management in Cloud Networking Using Economic Analysis and Pricing Models: A Survey

Nguyen Cong Luong, Ping Wang, *Senior Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Wen Yonggang, *Senior Member, IEEE*, and Zhu Han, *Fellow, IEEE*

Abstract—This paper presents a comprehensive literature review on applications of economic and pricing models for resource management in cloud networking. To achieve sustainable profit advantage, cost reduction, and flexibility in provisioning of cloud resources, resource management in cloud networking requires adaptive and robust designs to address many issues, e.g., resource allocation, bandwidth reservation, request allocation, and workload allocation. Economic and pricing models have received a lot of attention as they can lead to desirable performance in terms of social welfare, fairness, truthfulness, profit, user satisfaction, and resource utilization. This paper reviews applications of the economic and pricing models to develop adaptive algorithms and protocols for resource management in cloud networking. Besides, we survey a variety of incentive mechanisms using the pricing strategies in sharing resources in edge computing. In addition, we consider using pricing models in cloud-based Software Defined Wireless Networking (cloud-based SDWN). Finally, we highlight important challenges, open issues and future research directions of applying economic and pricing models to cloud networking.

Keywords- Cloud networking, resource management, pricing models, economic models.

I. INTRODUCTION

Cloud computing is becoming the platform of choice for a number of applications due to the advantages of high computing power, low service cost, high scalability, accessibility, and availability. Cloud computing is used as an integral part of society in various domains and disciplines such as education [1], commerce [2], health care services [3], transportation [4], and social networks [5]. Cloud computing is expected to bring huge new revenue opportunities. Recent reports have showed that the global revenue generated from cloud services is more than \$200 billion in 2016, and there will be approximately 3.6 billion Internet users accessing cloud services by 2018 (<http://www.statista.com>). On-demand services provided by cloud computing include, for example, Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). Google Docs [6], Google App Engine [7], and Amazon's Elastic Compute Cloud (Amazon's EC2) [8] are among the popular commercial services available in cloud computing.

The cloud computing infrastructure is typically hosted in data centers. Therefore, the current cloud services are based on parallel implementations in distributed data centers connected

with each other through high speed networks. For example, the EU-funded Scalable and Adaptive Internet Solution (EU-funded SAIL) project [9] investigates a combination of cloud computing infrastructure and networking capabilities, called *cloud networking*. Cloud networking actually considers the network beyond the data centers with the aim of providing both on-demand computing and network resources. With cloud networking, the resources and services can be provisioned from interconnecting distributed data centers owned by one or multiple providers, called *cloud data center networking*. The cloud resources and services can also be integrated with mobile networks, i.e., *mobile cloud networking*. Moreover, *edge computing* models are deployed in cloud networking to bring the cloud resources and services close to users, and thus minimize overall costs, jitter, latencies, and network load. The aforementioned models of cloud networking along with the integration of the Software-Defined Networking (SDN) technology are expected to support and satisfy a large number of users and applications in terms of flexibility, cost and availability of the services.

However, managing network and cloud resources together in cloud networking has many challenges. It is crucial to have an integrated view of the existing physical and virtual topologies and characteristics of the resources, as well as the status of all network entities. Besides, the provisioning and placement of virtual resources must be done in the best way possible, taking into account the available resources of both the cloud and networks. Moreover, the reconfigurations must often be performed to resize or release the existing virtual resources due to, e.g., the dynamic network environments (node or link failure), and the variability/elasticity of resource demand. Inefficient resource management negatively affects performance and cost as well as impairing system functionality.

To address the aforementioned challenges, it is vital to develop resource management approaches which guarantee the scalability, efficiency, manageability, adaptability and reliability for cloud networking. Traditional approaches, e.g., the system optimization, merely focus on the system performance metrics given system parameters and constraints rather than economic factors, e.g., the profit, cost, and revenue. Therefore, economic and pricing approaches have been recently explored, developed, and adopted for resource management in cloud networking. Compared with the system optimization approaches, the economic and pricing approaches provide the following advantages:

- In cloud networking, the profits of cloud providers have to be maximized while meeting the user demands. Thus, the

N. C. Luong, P. Wang, D. Niyato, and Y. Wen, are with School of Computer Science and Engineering, Nanyang Technological University, Singapore. E-mails: clnguyen@ntu.edu.sg, wangping@ntu.edu.sg, dniyato@ntu.edu.sg, and ygwen@ntu.edu.sg.

Z. Han is with Electrical and Computer Engineering and Computer Science, University of Houston, Houston, TX, USA. E-mail: hanzhu22@gmail.com.

profit guarantee for all cloud providers is a primary goal. Pricing models based on, e.g., the profit maximization or cost minimization, have been efficiently used to achieve the goal.

- There are various actors/stakeholders in cloud networking which belong to different entities, e.g., end-users, infrastructure providers, service providers, brokers, and network operators. They have different objectives, e.g., the profit, revenue, cost and utility, as well as different constraints, e.g., the budget and technology. Their objectives often conflict with each other, and this makes economic and pricing models become effective tools in cloud networking. More specifically, through the use of negotiation mechanisms, economic and pricing approaches can determine optimal solutions for selfish entities given their constraints.
- The demand for cloud computing and network resources depends on many users' attributes, e.g., the willingness to pay and performance requirements. Pricing strategies which rely on the demand elasticity such as price discrimination have been recently used as ideal solutions to optimize the provisioning of resources and profits of providers.
- Video on Demand (VoD) undoubtedly is among the most important services in cloud networking. Several commercial video delivery services have been introduced and become popular, e.g., YouTube and Netflix. However, the bandwidth cost of the service is typically very significant. Pricing mechanisms, e.g., smart data pricing, have been applied to regulate the user demands and maximize the bandwidth utilization.
- Besides the high bandwidth utilization for providers, guaranteeing quality of service (QoS), e.g., a small delay, for users is very important. Pricing approaches provide very efficient solutions for the joint optimization of both providers and users.
- To reduce service delay for users, cloud networking has developed edge computing models which employ devices at network edges to provide closer cloud resources and services to users. Pricing and payment strategies stimulate users to use the edge resources rather than distant data centers while still guaranteeing profits for cloud providers.

Although there are several surveys related to cloud networking, they do not focus on economic and pricing approaches, which are emerging as a promising tool. For example, a survey of applications of network virtualization for cloud computing was given in [10]. The survey of technologies of the Network-as-a-Service (NaaS) paradigm for supporting network-cloud convergence was presented in [11]. There are also surveys related to the architecture of SDNs, e.g., [12], [13], [14], [15], and applications of edge computing [16]. There are surveys related to the pricing approaches, e.g., [17], [18], [19]. However, they addressed the issues in Internet or wireless networks only. To the best of our knowledge, there is no survey specifically discussing the use of economic and pricing models to deal with resource management in cloud networking. This

motivates us to develop the survey with the comprehensive literature review on the economic and pricing models in cloud networking.

For convenience, the related works in this survey are classified based on various models of cloud networking and then their issues as shown in Table II. The models of cloud networking considered in this survey are cloud data center networking, mobile cloud networking, edge computing, and cloud-based Video-on-Demand (VoD) systems. Furthermore, some pricing approaches for the resource management in cloud-based Software Defined Wireless Networking (cloud-based SDWN) are discussed. Advantages and disadvantages of each approach are highlighted.

The rest of this paper is organized as follows. Section II describes a general architecture of cloud networking. Section III introduces the fundamentals of economic and pricing models. Section IV discusses how to apply economic and pricing models for resource management in cloud data center networking such as bandwidth, request, and workload allocation. Applications of economic and pricing models for resource allocation in mobile cloud networking are given in Section V. Section VI reviews economic and pricing models to address issues concerning the bandwidth allocation, task allocation, and storage sharing in edge computing. Section VII considers economic and pricing approaches for bandwidth allocation and Peer-to-Peer (P2P) caching in cloud-based VoD system. In addition, applications of economic and pricing models for bandwidth allocation and mobile data offloading in cloud-based SDWN are given in Section VIII. We outline important challenges, open issues, and future research directions in Section IX. Finally, we conclude the paper in Section X. The list of abbreviations appeared in this paper are given in Table I.

II. GENERAL ARCHITECTURE OF CLOUD NETWORKING

A. Definition of cloud networking

The term cloud networking is understood in a multi-administrative domain scenario in which network and data center domains interact with each other through predefined interfaces [20], [21]. Specifically, cloud networking extends network virtualization beyond the data centers to provide cloud and network resources to clients/users. Network resources can be virtual routers, bandwidth, virtual firewalls, or any network management software.

The definition also shows a key difference between the cloud networking and traditional computer networks, that is the *network virtualization*. By using network virtualization, the cloud networking reduces the cost for both providers and clients through real-time, on-demand resource and service provisioning. The resources are assigned and used by the client's needs, and the client only pays for what is used [22]. On the contrary, resource allocation in traditional computer networks is static, and a client needs to pay for every cost regardless of whether the resource has been used or not.

B. Architecture of cloud networking

The goal of a cloud networking architecture is to enable an efficient composition of cloud and network resources in a

TABLE I
MAJOR ABBREVIATIONS

Abbreviation	Description
BBU	BaseBand processing Units
Cloud-RAN	Cloud-Radio Access Network
CAPEX	CAPital EXpenditure
CWMSN	Cloud-based Wireless Multimedia Social Network
IaaS	Infrastructure-as-a-Service
IoT	Internet of Things
MCN	Mobile Cloud Networking
MNO	Mobile Network Operator
NFV	Network Function Virtualization
NUM	Network Utility Maximization
OPEX	OPerational EXpenditure
P2P	Peer-to-Peer
PaaS	Platform-as-a-Service
RRH	Remote Radio Head
SaaS	Software-as-a-Service
SDN	Software-Defined Networking
SDWN	Software-Defined Wireless Networking
SLA	Service-Level Agreement
SP	Service Provider
VCG	Vickrey-Clarke-Groves
VM	Virtual Machine
VoD	Video on Demand

cloud environment. To achieve the goal, several architectures were proposed for cloud networking. They can be based on intra-data center networking and inter-data center networking which are commonly called the cloud data center networking [11], [20], [23], [24], [25]. They can be based on mobile cloud networking [26], [27] or edge computing models [28] [29], [30], [31].

Based on these architectures, we provide a general, unified architecture for cloud networking as shown in Fig. 1. The architecture has three major parts: (i) cloud data center networking, (ii) mobile cloud networking, and (iii) edge computing. Their descriptions are given in what follows. Note that these parts can be independent from each other. Stakeholders or actors commonly participating in cloud networking are as follows.

- *Cloud provider*: A cloud provider, e.g, an IaaS cloud provider, owns and manages data centers and system software.
- *Network provider*: A network provider provides network connectivities among data centers of cloud providers or between end-users and data centers. In cloud networking, the network providers aim at cooperating with cloud providers to allocate cloud network resources and services to end-users or cloud users.
- *Cloud tenant/cloud user*: A cloud tenant can be a service provider, an organization or an enterprise, which uses cloud resources to host applications offered to its end-

users. Netflix (<https://www.netflix.com/>) is an example of a service provider of video on demand.

- *Cloud service broker*: A cloud service broker (or broker for the sake of shortness) acts as an intermediary between cloud users/end-users and cloud providers.
- *End-users*: The users generate resource and service requests or workloads that need to be processed using cloud resources.

1) *Cloud data center networking*: A data center is a large group of networked computer servers which are capable of providing the remote storage, processing, or distribution of large amounts of data. We provide brief descriptions of the components and resources in both intra- and inter-data center networking to which they will be referred in this survey.

- *Intra-data center networking*: Intra-data center networking refers to the interconnection between servers and storage resources through a networking system within a data center. The networking system includes virtual switches, Top-of-Rack (ToR) switches, core switches, and non-blocking switch.
 - *Virtual Machine (VM)*: VM is a software program or operating system which is able to perform tasks such as running applications and programs as a separate computer [32]. Multiple VMs can exist within a physical server or machine through virtualization techniques. In cloud networking, a VM can be migrated among servers within a data center or between data centers owned by different providers.
 - *Virtual switch*: A virtual switch is generally a software-based Ethernet switch function running inside a server. It can support Ethernet and/or IP services and provide switching and routing context separation among tenants/users sharing the same server.
 - *Network slicing*: Network slicing allows compartmentalizing VMs of the same application into the same virtual networks [33] and guarantees virtual resource isolation and virtual network performance.
 - *ToR switch*: A ToR switch supports Ethernet virtual LAN (VLAN) services or simple IP routing for the data center. The ToR switch aggregates Ethernet links from the servers. ToR switches are connected to one or two core switches in a data center.
 - *Non-blocking switch*: A switch is called non-blocking if it is able to connect all ports such that any routing request to any free output port can be established successfully without interfering other traffics.
 - *Core switch*: A core switch hosts multiple ToR switches and large-scale virtual LAN services or simple IP routing for the data center.
- *Inter-data center networking*: Data centers can be interconnected across the Wide Area Network (WAN) using inter-data center networking. Some commonly referred entities in the inter-data center networking are as follows.
 - *Data center gateway*: A data center gateway provides connectivity among data centers and to Internet and

TABLE II
A TAXONOMY OF THE APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD NETWORKING

System models Design issues	Cloud data center networking (Section IV)	Mobile cloud networking (Section V)	Edge computing (Section VI)	Cloud-based VoD system (Section VII)	Cloud-based SDWN (Section VIII)
Bandwidth allocation	✓	✓	✓	✓	✓
Resource allocation		✓	✓		
Task allocation			✓		
Request allocation	✓				
Workload allocation	✓				
Storage sharing			✓		
P2P caching				✓	
Mobile data offloading					✓

VPN customers. The data center gateway can provide virtual routing and switching capabilities.

- *IP/MPLS network*: An Internet Protocol/Multi-Protocol Label Switching (IP/MPLS) network is a packet-switched network that employs the Internet Protocol (TCP/IP) enhanced with the MPLS standard.
- *Resource pool*: A resource pool is a collective set of resources in data centers.
- *Federated cloud networking*: Federated or federation cloud networking refers to the cooperation among cloud providers to establish the federated cloud resource. For the federated cloud networking, a cloud provider can “borrow” cloud resources from other providers if its own resources are overloaded. This is called *outsourcing*. Also, a cloud provider can “rent out” its resources to other cloud providers if its resources are free. This is called *insourcing*.

2) *Mobile cloud networking*: Mobile Cloud Networking (MCN) is the EU FP7 Large-scale Integrating Project (IP) (cordis.europa.eu/fp7/ict/future-networks). It focuses on integrating the cloud computing and network function virtualization technologies to mobile networks [34]. MCN is able to provision services involving mobile network, decentralized computing, and storage as one on-demand unified service. The main characteristics of MCN are as follows [26]:

- MCN improves the real-time performance of mobile network functions, e.g., the baseband unit processing, mobility management, and QoS control, based on the high-performance cloud computing infrastructure. Thus, MCN enables adapting to the elasticity of the load.
- MCN provides an entirely new mobile cloud application platform as well as novel revenue streams for Telco by orchestrating infrastructure and services across different domains including wireless, mobile core networks, and data centers.
- MCN has the 3GPP LTE compliant architecture to exploit and support cloud computing.
- MCN introduces a new business actor, i.e., the MCN provider, in addition to typical stakeholders, e.g., the cloud computing provider, application provider, and

users.

A wide range of services is offered by MCN: (i) typical cloud computing atomic services, e.g., the computing, storage, and networking, (ii) support services, e.g., Monitoring as a Service (MaaS), (iii) virtualized network infrastructure services, e.g., Radio Access Network-as-a-Service (RANaaS) and Evolved Packet Core-as-a-Service (EPCaaS), (iv) new virtualized applications and services, e.g., Content Delivery Networks-as-a-Service (CDNaaS), and (v) End-to-End (E2E) services. In particular, RANaaS allows to partially move functionalities of RAN, i.e., digital processing functions, to a data center depending on the actual needs and network characteristics [35]. When all RAN functionalities are shifted towards the data center, and only RF functions are performed at Remote Radio Head (RRH), we have the concept of *Cloud-RAN* or *Centralized-RAN* [36]. The RANaaS implementation has the following major characteristics [37]:

- *On-demand provisioning*: Mobile network resources and services are provisioned according to the demand elasticity of mobile users.
- *Virtualization of RAN resources and functions*: They aim at optimizing usage, management, and scalability of the mobile network.
- *Resource pooling*: This allows virtual operators to share more dedicated resources and services, and thus enabling more business opportunities.
- *Elasticity*: This characteristic enables scaling network resources at the data centers or controlling the number of active RRHs.
- *Service metering*: Operators provision and charge RAN operation services, e.g., the usage of RRHs, on a measurable and controllable basis.
- *Multi-tenancy*: This feature ensures the security in the mobile network by enabling isolation mechanisms and charging of different users.

3) *Edge computing*: Edge computing is a paradigm which pushes the frontier of computing applications, data, and services away from central nodes, e.g., the data centers, to the periphery or edges of the network [38]. Edge computing covers a wide range of technologies including cloudlet, remote/micro/community clouds, nano data centers, volunteer

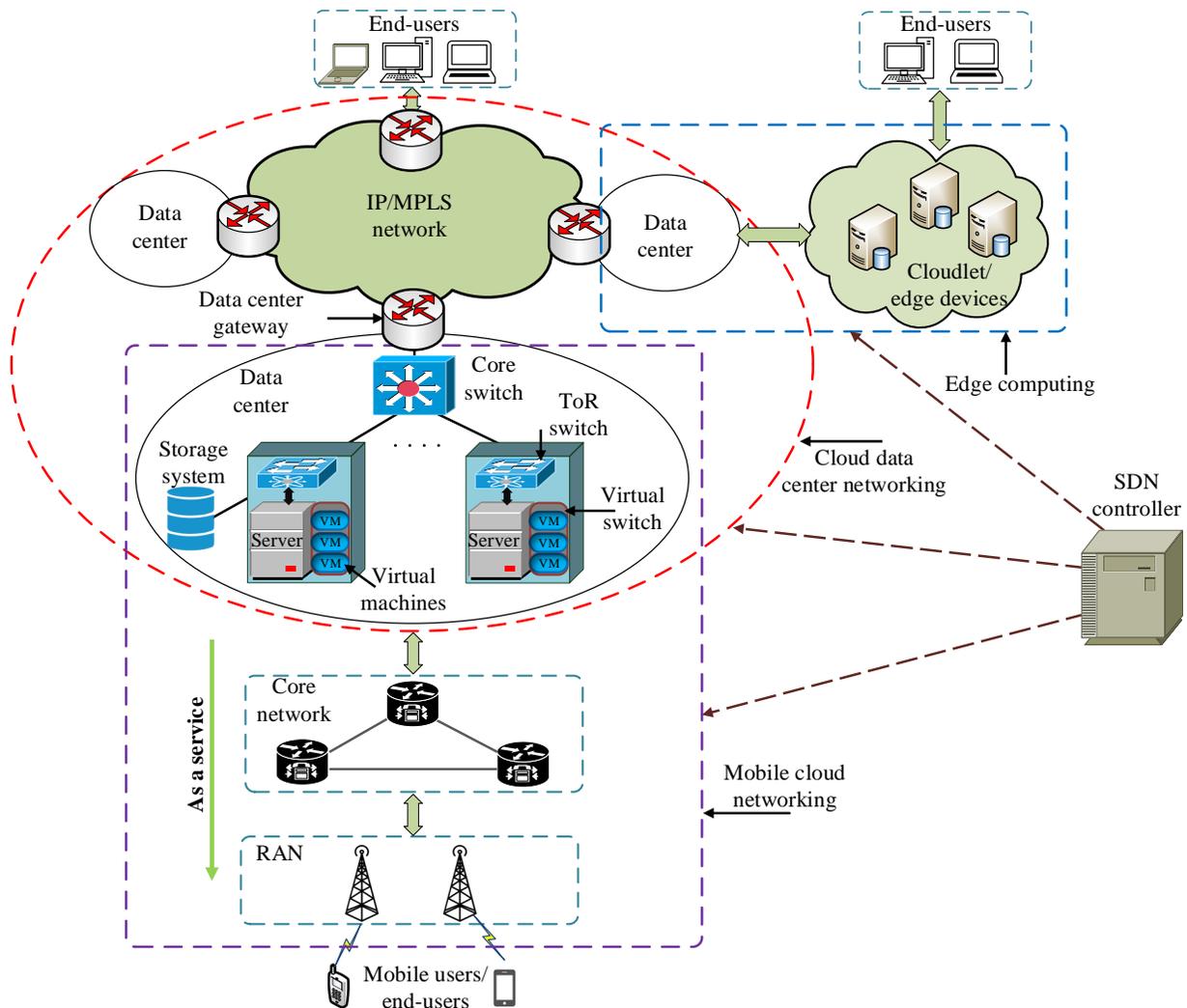


Fig. 1. A general architecture of cloud networking.

computing system, local cloud/fog computing, client-assisted cloud system, sensing networks, e.g., the wireless sensor network and crowdsensing network, and distributed Peer-to-Peer (P2P). Edge computing has the following major advantages [39]:

- It significantly reduces the data traffic, cost, and latency and improves QoS since cloud resources and services are located close to users.
- It alleviates the major bottleneck and the risk of a potential point of failure since it does not rely on centralized computing.
- It enhances security since data is encrypted as the data is moved towards the network edge.
- It provides high levels of scalability, reliability, and automation.

4) *Software-Defined Networking (SDN)*: A traditional network architecture is composed of three planes of functionality, i.e., data, control and management planes. The control plane makes forwarding/routing decisions on the user traffic based on forwarding/routing tables. The data plane is responsible for forwarding the user traffic using the decisions from the

control plane. The control and data planes are always coupled and embedded in the same networking devices, e.g., switches and routers, to guarantee network resilience. However, such architecture is rigid and complex to manage and control [40], [41]. The Software-Defined Networking (SDN) [13], [42], [43] is an emerging networking paradigm towards simple and flexible network management for network operators. SDN is defined as “a network architecture where the forwarding state in the data plane is managed by a remotely controlled plane decoupled from the former” [13]. In other words, SDN decouples the control plane from the network devices to become an external entity, the so-called *SDN controller*. The SDN architecture has four major features below:

- The control and data planes are decoupled, and network devices just act as forwarding elements.
- Forwarding decisions are flow-based instead of destination-based, meaning that all packets in the same flow receive identical service policies at the forwarding devices. This allows to unify different types of network devices, e.g., routers, switches, firewalls, load-balancers, and traffic shapers.

- The control logic is moved to an external entity, i.e., the SDN controller or the Network Operating System (NOS).
- The network is programmable through software applications running on the SDN controller which interacts with the underlying data plane devices.

These features of SDN make the networks more programmable and easily partitionable and virtualizable. In practice, SDN has been used to address many issues in a wide range of network environments [44]. For example, it was used to address the security and resource allocation in enterprise networks [45], [46], flow control, virtual data center embedding, and resource utilization maximization in cloud networking [24], [47], [48], mobility management and load balancing in wireless access networks [49], wavelength path control and QoS-aware unified control in optical networks [50], and network management in home and small business networks [51]. In particular for the cloud networking, using SDN makes a number of network devices become simple forwarding elements which are cheap and easy to deploy. This reduces both capital and operational expenditures for cloud and service providers. It is also expected to significantly improve benefits for all stakeholders, especially when SDN can be combined with the economic and pricing models which will be discussed in the next section.

III. OVERVIEW AND FUNDAMENTALS OF ECONOMIC AND PRICING THEORY APPLIED IN CLOUD NETWORKING

Economic and pricing approaches have been applied to address many issues in cloud networking due to the aforementioned benefits. In this section, we classify the economic and pricing approaches commonly used for resource management in cloud networking as shown in Fig. 2. The classification is based on how the prices are set, i.e., market-based pricing, game theoretic and auction based pricing, and Network Utility Maximization (NUM) based pricing.

A. Market-Based Pricing

In the following, we present the pricing models based on economic and financial concepts which have been applied in the cloud networking. We first present a simple pricing model, i.e., cost-based pricing, and then describe more complex pricing models including differential pricing, profit maximization pricing, and Ramsey pricing.

1) *Cost-based pricing*: Cost-based pricing is a common pricing strategy to determine the price of a service based on calculating the total cost of the service and adding a percentage of the cost as a desired profit. The objective of using the cost-based pricing is to ensure that the price makes the service provider profitable, or at least the price covers the total cost of the service provider. The total cost generally consists of a fixed cost and a variable cost. The fixed cost is the cost that does not change when the number of sales of the services changes. For example, hardware costs, e.g., servers and network devices. On the contrary, the variable cost varies according to the number of sales of the services produced. For example, the resource costs, e.g., energy and bandwidth costs, the cost of data transfer between different data centers, and the

cost of cloud server usage, are the variable cost for generating cloud services. The advantage of the cost-based pricing is the ease of setting the price since the price is a function of the internal cost, i.e., the cost required to generate the service [52]. However, this pricing strategy does not consider external market factors, e.g., the pricing strategies of other providers and the perceived value and willingness to pay of buyers.

In cloud networking, the cost-based pricing has been used by cloud providers for evaluating the service cost in geodiverse data center networks [53], [54]. It has been also employed to analyze the cost saving when SDN and the Network Function Virtualization (NFV) in the cloud are enabled [55], [56]. However, the internal cost information in cloud networking may not be easy to obtain due to the variable cost diversity. For example, the variable cost could depend on the geography of data centers. Further details on the cost-based pricing model can be found in [57], [58].

2) *Differential pricing*: The cost-based pricing ignores the requirements and preferences of cloud users or tenants. To maximize the profit of providers, differential pricing, also called price discrimination, can be used. Consider a cloud resource market consisting of a cloud provider with its cloud resources, i.e., the computing resource and network bandwidth. Using the differential pricing, the cloud provider may charge different prices to different cloud users based on their demand and willingness to pay. By setting higher prices for one type of user, the use of the differential pricing actually transfers the user surplus to the provider. Here, the user surplus is the difference between the total money that users are willing to pay and the total money that they actually pay. Thus, although this pricing guarantees a high revenue for the provider, it can be unfair to cause one type of user to pay a greater price than another type of user. In cloud networking, the differential pricing has been applied for bandwidth allocation among groups of users having different elasticities on cloud resources or among cloud users having different flexibility in resource usage as proposed in [59]. In current cloud service markets, the differential pricing is used to set prices of the cloud services based on the requirements of the users. For example, the Alibaba group (<https://intl.aliyun.com/>) offers lower prices to users which require the cloud services for long term, e.g., 1 year.

3) *Profit maximization*: Profit maximization is the process of determining the output quantity and the corresponding price which yield the highest profit for a provider [60]. We present briefly how to find the optimal quantity and price based on the profit maximization in the following. Assume that a cloud provider needs to determine the number of cloud resource (i.e., the computing and network bandwidth) units, denoted by Q and the corresponding price P for their cloud users. The profit of the cloud provider is $\pi = R(P, Q) - C(Q)$, where $R(\cdot, \cdot)$ is the total revenue and $C(\cdot)$ is the total cost. The total cost may involve a fixed cost and a variable cost. The revenue is the amount of money that the cloud provider receives from selling Q resource units to its users. The optimal quantity of cloud resource units, i.e., Q^* , is determined such that the profit is maximized, i.e., $Q^* = \max_Q \pi$. The optimal

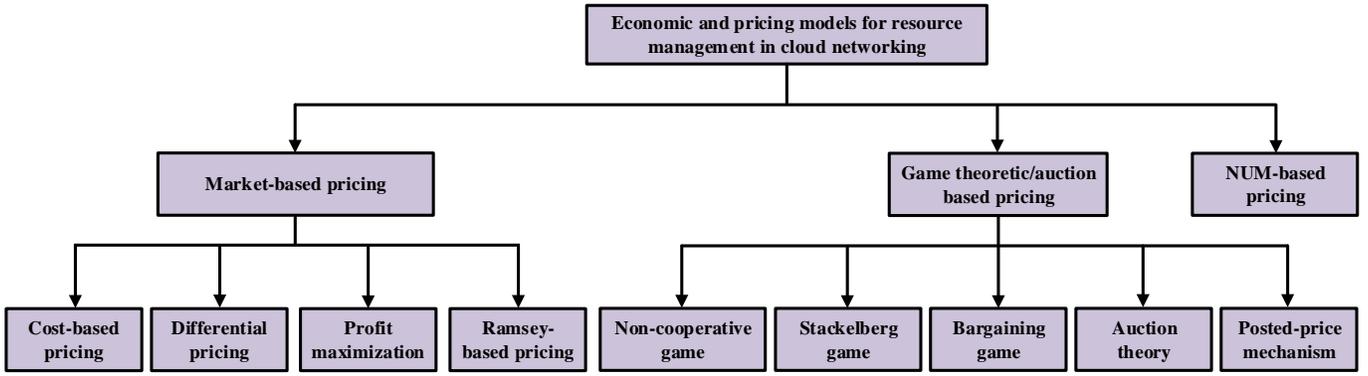


Fig. 2. A taxonomy of economic and pricing models in cloud networking.

quantity allows to find the optimal price based on the demand curve. The demand curve is typically a linear curve to show the relationship between the price of a resource unit and the quantity of resource units that users are willing to buy. A general demand curve can be expressed as $P = a - bQ$, where a and b are proper parameters. Thus, at $Q = Q^*$, the optimal price is $P^* = a - bQ^*$.

The optimal quantity and price can be determined using the graph as shown in Fig. 3(a) and Fig. 3(b). Fig. 3(a) shows the curves of the total cost, total revenue, and profit. The optimal quantity Q^* is determined at the positive peak value of the profit curve. Then, the optimal price P^* is obtained from the demand curve in Fig. 3(b). However, in real markets, it is not easy to determine the demand curve. The user demand can be random or assumed to follow some distributions, e.g., the Gaussian [61]. Moreover, the profit maximization does not consider the market competition in determining the quantity and price. In cloud networking, the profit maximization has been adopted to allocate computing and network resources to users [62] or to assign resource requests from users to cloud providers [63].

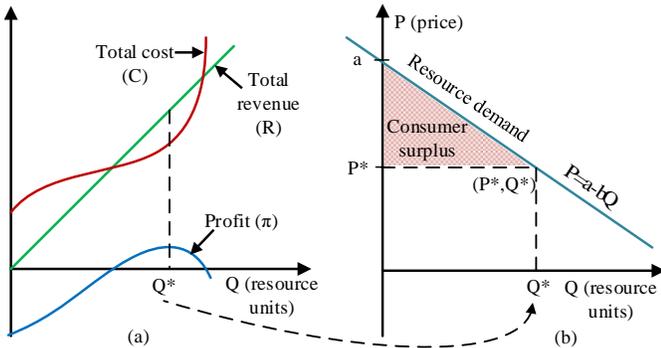


Fig. 3. Pricing based on profit maximization.

4) *Ramsey pricing*: In Ramsey pricing, different prices of the same commodity are applied to different markets depending on the demand elasticity of the commodity [64]. Ramsey pricing is similar to the differential pricing. However, unlike the differential pricing that maximizes the profit of the provider, Ramsey pricing aims to maximize the social welfare of users subject to a predefined threshold on the

provider's profit. Specifically, assume that a cloud provider provides cloud resource units in two independent markets. The cloud provider determines different prices (p_1, p_2) in the two markets. In the independent market setting, the demands of the resource corresponding to two prices are $(q_1(p_1), q_2(p_2))$. The marginal cost of offering one cloud resource unit in both markets is c , and the cloud provider has a fixed cost. The objective of the cloud provider is to determine (p_1, p_2) to maximize social welfare subject to the constraint that the profit of the cloud provider is not less than a threshold. Here, the social welfare is the consumer surplus which is the area on the left of the demand curve and above the price as shown in Fig. 3(b). The threshold is a fixed profit value π^* which is predefined by the cloud provider. Therefore, the optimization problem of the cloud provider can be formulated as follows:

$$\begin{aligned} \max_{(p_1, p_2)} \sum_{i=1}^2 \left(\int_{p_i}^{\infty} q_i(p) dp \right) \\ \text{s.t. } \sum_{i=1}^2 (p_i - c) q_i(p_i) \geq \pi^*. \end{aligned} \quad (1)$$

The problem in (1) can be solved using the Lagrange multiplier method. The relationship between the optimal price and the demand elasticity in market i , $i \in \{1, 2\}$, is expressed as follows:

$$\frac{p_i^* - c}{p_i^*} = \frac{1 - \lambda^*}{\lambda^*} \frac{1}{\epsilon_i}, \quad (2)$$

where λ is the Lagrange multiplier, and $\epsilon_i = \frac{dq_i(p_i^*)}{dp_i} \frac{p_i^*}{q_i}$ is the coefficient of price elasticity of demand for the resource in market i . The price elasticity of demand gives the percentage change in quantity demanded according to a fall or a rise in its price. We get the following relationship from (2):

$$\frac{p_1^* - c}{p_1^*} / \frac{p_2^* - c}{p_2^*} = \frac{\epsilon_2}{\epsilon_1}. \quad (3)$$

The expression in (3) shows that the price of a resource in a market should be relatively low when the demand elasticity in the market is high. On the contrary, in a market with inelastic or less elastic demand, the provider can set a high price because the demand for the resource does not change significantly according to a fall or a rise in the price. Clearly,

the provider desires to have a market with inelastic demand since it can increase the price to earn more revenue. However, determining demand elasticity in markets is challenging. Moreover, the cloud provider cannot apply this pricing in the long run since users charged with a higher price will seek alternatives [65]. One application of the Ramsey pricing in cloud networking is to regulate traffic flows of users among data centers [66], which will be discussed in Section IV-A8.

B. Game Theory and Auction Based Pricing

Game theory and auctions are the study of multiparticipant decision making problems in which a choice of a participant, i.e., a player, potentially affects the interests of other participants [67]. In the context of cloud networking, participants can be cloud providers, service providers, cloud tenants, and users. In the following, we briefly present game theoretic models and auction mechanisms which have been widely used to determine resource prices in the cloud networking. First, some important terminologies are defined below [68].

- *Player*: A player is a participant which makes a decision in the game.
- *Payoff*: A payoff, i.e., a utility, a profit, or an interest, reflects the desired outcome of the player.
- *Strategy*: Player's strategy is a set of actions/instructions that the player can follow to achieve a desired outcome. The payoff depends on not only the player's own action, but also the actions of others.
- *Rationality*: A player is rational if its strategy always aims at maximizing its own payoff.

1) *Non-cooperative game*: In the non-cooperative game, each player maximizes only its own payoff neither being concerned about the payoff of the other players nor about the social welfare of the network [69]. In this game, the players are selfish, and they do not form coalitions or make agreements with each other.

Consider a cloud resource market in which cloud providers as the sellers compete for selling resources to users. The sellers are typically selfish, and therefore the market can be modeled as a non-cooperative game among the sellers along with their pricing strategies. Assume that there are N players, and P_i is a set of pricing strategies of player i , where $P = P_1 \times \dots \times P_N$ is the Cartesian product of the individual strategy sets. Let $p_i \in P_i$ be the pricing strategy of player i . A vector of strategies of N players is $\mathbf{p} = (p_1, \dots, p_N)$, and a vector of corresponding payoffs is $\boldsymbol{\pi} = (\pi_1(\mathbf{p}), \dots, \pi_N(\mathbf{p})) \in R^N$, where $\pi_i(\mathbf{p})$ is the payoff of player i given the player's chosen strategy and strategies of the others. Each player chooses its best strategy p_i^* which maximizes its payoff. A set of strategies $\mathbf{p}^* = (p_1^*, \dots, p_N^*) \in P$ is the Nash equilibrium if no player can gain higher payoff by changing its own strategy when the strategies of the others remain the same [70], i.e.,

$$\forall i, p_i \in P_i : \pi_i(p_i^*, \bar{\mathbf{p}}_i^*) \geq \pi_i(p_i, \bar{\mathbf{p}}_i^*), \quad (4)$$

where $\bar{\mathbf{p}}_i = (p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_N)$ is a vector of strategy choices of all players except player i .

The inequality in (4) shows the stable state of the game in which the players have no incentive to change their own

strategies since the payoffs will be worse off. However, in some cases, there is no Nash equilibrium at all, or there may exist multiple Nash equilibria which can make players not be clear about which one to choose. Therefore, checking the existence and uniqueness of the Nash equilibrium is important when setting prices based on the non-cooperative game.

The non-cooperative game theory has been widely used for the resource management in cloud networking. For example, it has been used to model the bandwidth sharing among peers in cloud-assisted P2P streaming systems [71] or among brokers in cloudlet systems [72]. It has been adopted to maximize the profits of cloud providers as presented in [73].

2) *Stackelberg game*: The non-cooperative game discussed above assumes that players announce their pricing strategies simultaneously, and the players know each other's strategies at the same time. However, this may not always hold in real markets. Therefore, sequential games can be used in which players can announce their strategies following a certain predefined order. This is the Stackelberg game [74]. In the Stackelberg game, the player decides its own strategic choice after observing the strategies of other players [75]. It was proved that even if the players have to choose their strategies first, their payoffs are not less than those at the Nash equilibrium [76], i.e., due to the first-mover advantage. The following provides the definition and properties of the Stackelberg game.

Assume that there are two cloud resource sellers 1 and 2 in the market. P_1 and P_2 are the sets of pricing strategies of sellers 1 and 2, respectively. Seller 1 chooses its pricing strategy p_1 from set P_1 to maximize its payoff or profit function $\pi_1(p_1, p_2)$, and seller 2 chooses its pricing strategy p_2 from set P_2 to maximize its payoff function $\pi_2(p_1, p_2)$. Without loss of generality, assume that seller 2 selects its strategy before seller 1 decides its selection. Seller 2 is namely the leader, and seller 1 is called the follower. We have the following definition [77]:

Definition 1. If there exists a mapping $F : P_2 \rightarrow P_1$ such that, for any fixed $p_2 \in P_2$, $\pi_1(Fp_2, p_2) \geq \pi_1(p_1, p_2)$, $\forall p_1 \in P_1$, and if there exists $p_{2s2} \in P_2$ such that $\pi_2(Fp_{2s2}, p_{2s2}) \geq \pi_2(Fp_2, p_2)$, then the pair $(p_{1s2}, p_{2s2}) \in P_1 \times P_2$, where $p_{1s2} = Fp_{2s2}$, is called a Stackelberg strategy pair.

Definition 1 means that the Stackelberg strategy is optimal for the leader when the follower responds to the leader with the follower's optimal strategy. Let $D_1 = \{(p_1, p_2) \in P_1 \times P_2 : p_1 = Fp_2\}$ denote the rational reaction set of seller 1 when seller 2 chooses strategy $p_2 \in P_2$. Seller 1 is referred to as a rational player. Similarly, when seller 1 is the leader, let D_2 denote the rational reaction set of seller 2. The sets D_1 and D_2 have significant importance which is indicated in the following two propositions.

Proposition 1. A strategy pair (p_{1s2}, p_{2s2}) is the Stackelberg strategy with seller 2 as the leader iff $(p_{1s2}, p_{2s2}) \in D_1$ and

$$\pi_2(p_{1s2}, p_{2s2}) \geq \pi_2(p_1, p_2), \forall (p_1, p_2) \in D_1. \quad (5)$$

Proposition 2. A strategy pair (p_{1N}, p_{2N}) is the Nash strategy pair iff $(p_{1N}, p_{2N}) \in D_1 \cap D_2$.

The expression in (5) and Proposition 2 show that $\pi_2(p_{1s2}, p_{2s2}) \geq \pi_2(p_{1N}, p_{2N})$. In other words, for the leader,

the Stackelberg strategy guarantees to achieve the payoff at least as good as the corresponding Nash equilibrium. This is because when choosing the Stackelberg strategy, the leader actually imposes a solution which will be favorable to itself.

In cloud networking, the Stackelberg game has been applied for allocating the cloud provider's bandwidth to virtual networks [33] and for reducing the access of users to servers in the cloud [78]. The Stackelberg game has been also applied in cloud computing. For example, it was used to maximize revenue of the cloud provider while maximizing server clients' utilities [79] or to maximize revenue of the cloud provider while guaranteeing QoS for its end-users [80]. Besides, the Stackelberg game has been used in Internet of Things (IoT). For example, it was adopted to maximize the profits of different participants of IoT industry value chain [81] and to improve the QoS and the network's robustness in sensing networks [82].

3) *Bargaining game*: In the bargaining game or Nash bargaining game, two or more players must reach an agreement regarding how to distribute a monetary amount. Consider trading bandwidth in cloud networking between a cloud provider, i.e., a seller, and a cloud tenant, i.e., a buyer. A successful bargain is reached if and only if the bandwidth is allocated at a mutually acceptable price. Let p_s^0 be the smallest price that the seller can accept for selling the bandwidth and p_b^0 be the buyer's greatest price that the buyer is willing to pay for the bandwidth. The pair (p_s^0, p_b^0) is called the disagreement point or threat point that the seller and the buyer expect to receive if their negotiations fail to reach a settlement [83].

The strategy of the seller is to offer the selling price p_s^* to maximize its expected profit $\pi_s(p_s, p_s^0)$, i.e., $\pi_s(p_s^*, p_s^0) \geq \pi_s(p_s, p_s^0), \forall p_s$. Similarly, the strategy of the buyer is to offer the buying price p_b^* to maximize its profit $\pi_b(p_b, p_b^0)$, i.e., $\pi_b(p_b^*, p_b^0) \geq \pi_b(p_b, p_b^0), \forall p_b$. If $p_b^* \geq p_s^*$, a bargain is enacted and the transaction price for trading the bandwidth can be set by [84], $p^* = kp_b^* + (1 - k)p_s^*$, with $0 \leq k \leq 1$. When $k = 1/2$, the transaction price is determined by splitting the difference between the buyer's and seller's offers. A pair of the best response offer strategies (p_s^*, p_b^*) constitutes the Nash bargaining solution. At this agreement point, the seller earns $(p^* - p_s^0)$, and the buyer earns $(p_b^0 - p^*)$.

Some other scenarios in cloud networking where the bargaining game has been applied are allocating requests of users to data centers [85] and sharing cloud resources among service providers [86]. In cloud computing, the bargaining game has been used for negotiating the price among the cloud resource brokers and grid service providers as proposed in [87] and for pricing and allocating virtual resource instances for independent tasks and workflow tasks as presented in [88].

4) *Auction*: An auction is the economic mechanism the goals of which are to allocate commodities and establish corresponding prices via a process known as bidding [89]. There are some common terminologies used in the auction as follows:

- *Bidder*: A bidder is a buyer which wants to purchase resources. In cloud networking, bidders can be end-users or cloud tenants.

- *Seller*: A seller, e.g., a cloud provider, offers its resources and services for sale.
- *Auctioneer*: An auctioneer acts as an intermediate agent to conduct an auction, determine, and announce the winner. In many cases, an auctioneer is a seller itself.
- *Price*: A price in an auction may be a bidding price or an asking price. The bidding price is the price that the bidder is willing to pay for a requested resource, and the asking price is the price of a resource that the seller is willing to offer.

There exist several studies on auctions as well as their applications. There are a survey of the auction theory [90], a survey of auction on Internet [91], or a survey of auction approaches for resource management in wireless networks [92]. In what follows, we discuss typical types of auctions which have been commonly applied to resource management in cloud networking.

(a) *Conventional auctions*: A conventional auction is known as the *open-outcry* auction. In the open-outcry auction, bids of buyers are disclosed to each other during the auction. There are two types of the conventional auction [93].

- *English auction*: The English auction is an ascending-bid auction, meaning that the bidding price submitted by buyers increases monotonically. Specifically, buyers submit their bidding prices for the resource sequentially or simultaneously to the auctioneer. The auction will terminate if there is no new higher price submitted. The buyer with the highest price wins the resource and pays the price p^* , i.e., a hammer price, which satisfies $p_s^0 \leq p^* \leq \max_i B_i$, where p_s^0 is the lowest price that the seller can accept to sell, and B_i is buyer i 's budget. Generally, p^* changes depending on the number of buyers in the auction and may not equal p_s^0 .
- *Dutch auction*: Contrary to the English auction, the Dutch auction is a descending-bid auction in which the auctioneer or seller initially sets a high asking price for the resource and then decreases the price until one of the buyers accepts the price. The winning buyer pays the final price and receives the resource. This simple allocation enables the Dutch auction to spend less time than the English auction [94].

(b) *Vickrey and Vickrey-Clarke-Groves (VCG) auctions*: Vickrey and VCG auctions are the sealed-bid auctions in which buyers submit simultaneously their sealed bids to the auctioneer. Different from the open-outcry auctions, in the sealed-bid auction, buyers do not know bidding strategies of each other and cannot change their own bids during the auction.

- *Vickrey auction*: A Vickrey auction, also known as the second-price sealed-bid auction, is one of the two most important k -th-price sealed-bid auctions. In the Vickrey auction, the winning buyer pays the second-highest price rather than the price that it submitted [95], i.e., $p^* = \max_{p \in P \setminus \{p_i\}} p$, where p_i is the highest price of the winner. In other words, the winner pays a price less than its expected price [96]. Therefore, the Vickrey auction motivates buyers to bid truthfully, and such an auction achieves strategy-proofness, or incentive compatibility, or

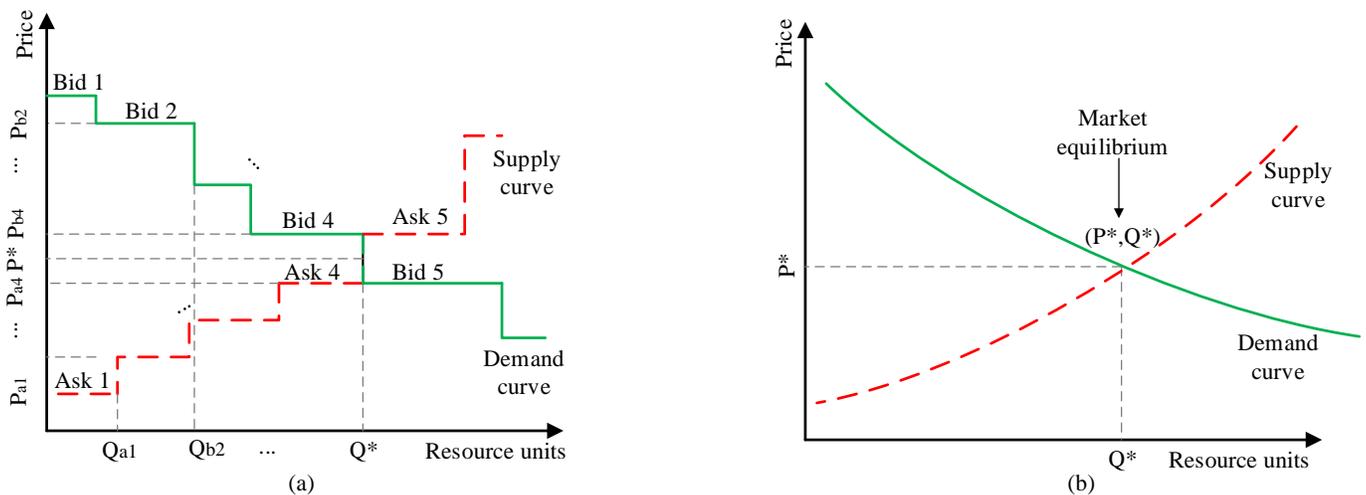


Fig. 4. (a) Discrete supply and demand curves of double auction, and (b) continuous supply and demand curves from economics.

truthfulness. The truthfulness is an important property because an auction which does not hold this property may be vulnerable to market manipulation and produce very poor outcomes [97].

- *Vickrey-Clarke-Groves (VCG) auction*: The VCG auction is a generalization of the Vickrey auction with multiple commodities [98]. The VCG auction allocates commodities in a socially optimal manner and charges the winner with the loss of the social value due to its getting the commodity.
 - Assume that there is a set T of M commodities for sale $T = \{t_1, t_2, \dots, t_M\}$, where t_i is the i th commodity, and a set of N buyers, i.e., bidders, $B = \{1, 2, \dots, N\}$.
 - Let $b_i(t_j)$ denote a bid of bidder i for commodity t_j and V_N^M denote the social welfare value, i.e., the social cost, created by M commodities.

Similar to the Vickrey auction, if $b_i(t_j)$ is the highest, bidder i wins to obtain commodity t_j . According to the VCG auction rule, bidder i pays the price which is equal to

$$V_{N \setminus \{i\}}^M - V_{N \setminus \{i\}}^{M \setminus \{t_j\}}, \quad (6)$$

where $V_{N \setminus \{i\}}^M$ represents the social welfare value if bidder i does not participate in the auction, and $V_{N \setminus \{i\}}^{M \setminus \{t_j\}}$ is the attainable social welfare value after bidder i wins commodity t_j . The expression in (6) indicates that winner i needs to pay the price which is the loss in attainable welfare suffered by the remaining bidders since it got the commodity t_j .

(c) *Forward, reverse, and double auctions*: Considering the sides of sellers and buyers, we can classify auctions as follows:

- *Forward and reverse auctions*: In the forward auction, there are many potential buyers and one seller. On the contrary, in the reserve auction, there are one buyer and many potential sellers.
- *Double auction*: In the double auction, buyers and sellers simultaneously submit their bids and asks to an auctioneer, respectively [99]. The basic idea of the double

auction is to match asks from sellers and bids from buyers by assigning commodities from the sellers to the buyers and payments from the buyers to the sellers accordingly. To further understand the double auction process, we consider a cloud resource market in which the sellers are cloud providers and the buyers are cloud tenants/users. Upon receiving bids and asks, the auctioneer sorts the buyers' bids in a non-ascending order and the sellers' asks in a non-descending order. The auctioneer finds the largest index k at which the asking price is less than the bidding price, i.e., $p_k^a \leq p_k^b$. The transaction price p^* , i.e., a hammer price or clearing price, can be determined as $p^* = (p_k^a + p_k^b)/2$. The buyer receives resources, and the seller gets payment p^* . The process is repeated to match the remaining buyers and sellers as well as to determine corresponding clearing prices.

In fact, the double auction has a very similar concept to the supply and demand model [100]. As shown in Fig. 4(a), the asks from sellers and the bids from buyers form the supply and demand curves, respectively. The x-axis represents the supplied resource units, and y-axis represents the asking or the bidding prices. For example, seller 1 sells Q_{a1} units of resources at price P_{a1} (i.e., "Ask 1" in Fig. 4(a)), and buyer 2 bids to buy Q_{b2} units of resources at price P_{b2} (i.e., "Bid 2" in Fig. 4(a)), and so on. Fig. 4(a) is actually a discretized form of the standard supply and demand model which is shown in Fig. 4(b). The supply and demand curves intersect at a point which is called the supply-demand equilibrium [101], i.e., the market equilibrium point (P^*, Q^*) . The clearing price P^* at the equilibrium can be determined by $(P_{a4} + P_{b4})/2$ (as shown in Fig. 4(a)).

The double auction can hold important properties: individual rationality (i.e., no participant loses when joining the auction), balanced budget (i.e., the auctioneer gains money), truthfulness (i.e., buyers and sellers submit truthfully their bids and asks), and economic efficiency (i.e., the social welfare is the best possible).

(d) *Combinatorial auction*: In the combinatorial auction, each bid of a buyer indicates a combination or package of discrete commodities rather than an individual commodity [102]. Given a set of bids, the auctioneer finds an optimal allocation of the commodities to the buyer, i.e., the winner. The combinatorial auction has more advantages compared with standard auctions, e.g., the sealed-bid auction. These advantages include little global information requirement, economic efficiency, utility maximization for buyers, and revenue maximization for sellers. However, a big challenge in the combinatorial auction is the winner determination problem. This problem is NP-hard, and there does not exist a polynomial-time algorithm to find the optimal allocation. However, many algorithms have been proposed to find approximate solutions for the problem, e.g., the Lagrangian relaxation approach [103].

(e) *Shapley value*: The aforementioned auction schemes only consider how to determine the winner and payment. In practice, the winner can be a group of users, which is called a *virtual winner* [104]. How to fairly allocate resources and share the payment among users in the virtual winner is still an important problem. The Shapley value method can be combined with an auction to solve the problem. This section provides the definition, properties, and applications of the Shapley value.

The Shapley value is a concept in the cooperative game theory which provides a unique and fair allocation/distribution of the total surplus/profit generated by the coalition among players [105]. Specifically, in a coalition, players may contribute differently to obtain an overall surplus/profit. The Shapley value provides a method to measure the importance of each player in the cooperation and to assign the distribution of generated surplus among the players. Formally, assume that there is a coalitional game (v, \mathbb{N}) , where \mathbb{N} denotes the coalition of $|\mathbb{N}|$ players, and v or $v(\mathbb{N})$ describes the total expected surplus obtained from the cooperation of the $|\mathbb{N}|$ players. The amount of surplus that player i in the coalitional game receives is defined by

$$\phi_i(v) = \sum_{\mathbb{S} \subseteq \mathbb{N} \setminus i} \frac{|\mathbb{S}|!(|\mathbb{N}| - |\mathbb{S}| - 1)!}{|\mathbb{N}|!} (v(\mathbb{S} \cup i) - v(\mathbb{S})), \quad (7)$$

where $v(\mathbb{S})$ is the total surplus obtained from the cooperation of $|\mathbb{S}|$ players in a coalition \mathbb{S} . The expression in (7) means that to calculate the Shapley value of player i in the coalition game (\mathbb{N}, v) , we need to determine the marginal contribution of the player in a potential coalition (\mathbb{S}, v) as $v(\mathbb{S} \cup i) - v(\mathbb{S})$ that is the difference between the total surplus of the game with player i and the game without player i . We then take the average of this contribution over all possible different permutations in which the coalition can be formed. The Shapley value satisfies the following axioms [106], [107]:

- *Joint efficiency*: The total surplus of all players equals that of the coalition, i.e., $\sum_{i \in \mathbb{N}} \phi_i(v) = v(\mathbb{N})$.
- *Zero payoff to the dummy*: A dummy player is the one which does not contribute anything to the value of any coalition that it joins, i.e., $v(\mathbb{S} \cup i) = v(\mathbb{S})$, for all \mathbb{S} . The payoff to the dummy player is zero.
- *Symmetry*: If two players have the same contribution, i.e.,

$v(\mathbb{S} \cup i) = v(\mathbb{S} \cup j)$, they receive the same payoff.

- *Additivity*: The payoff of any player is equal to the sum of all the payoffs that the player will receive as a member of all possible coalitions.

It was shown in [108], [109] that the Shapley value given in (7) is unique satisfying the four above axioms. This is desirable from the perspective of cooperative providers. In cloud networking, the Shapley value has been used for sharing profit among cloud providers from their resource pooling. Moreover, it has been also adopted for sharing the cost among cloud users from using bandwidth [110] and for the fair profit sharing among Internet service providers [111].

The summary of the above auctions along with their applications in cloud networking is given in Table III. As seen, auction mechanisms have been used for resource management in cloud networking. Moreover, the Vickrey auction, i.e., the second-price sealed-bid auction, has been more frequently used compared with the other auction mechanisms due to its privacy and truthfulness guarantee [112].

5) *Posted-price mechanism*: The posted-price mechanism is typically used in online procurement markets, e.g., a digital market, in which sellers arrive in a sequential order. The posted-price mechanism assigns a specific price to each seller when the seller arrives. The seller can “take or leave” this price. Typically, the seller accepts the price if its actual cost is less than the price offered by the mechanism. The seller then gives its buyers “take-it-or-leave-it” offer price, meaning that a buyer can accept or reject the offer [113]. Based on the buyer’s responses, the mechanism will set prices for subsequent sellers’ commodities. Since the posted-price mechanism may set different prices for the same commodity, it is similar to the differential pricing mechanism, i.e., the price discrimination, as presented in Section III-A2. Thus, the term “posted-price discrimination” sometimes appears in the literature [114]. Besides, making a comparison with the auction mechanisms, the posted-price mechanism is less complex since it decides immediately an offer price without soliciting an ask from each seller upon the seller’s arrival [115]. In the context of cloud networking, the posted-price mechanism has been used when resource sellers, e.g., users in the social cloud, arrive in a sequential order to offer their storage services. In commercial clouds, e.g., Amazon’s *elastic cloud compute (EC2)*, the cloud provider uses this mechanism to post a certain price on the “take-it-or-leave-it” basis [116]. Extended to IoT, this mechanism has been also used for the data aggregation in crowdsensing networks in which the data sellers are phone users [117], [118].

C. Network Utility Maximization (NUM)-based pricing

In this section, we explain the Network Utility Maximization (NUM), known as a dual-based distributed algorithm for the resource allocation. NUM is essentially the problem of maximizing the total utility of users in a network, given the capacity constraint of the network [119]. The original NUM problem only considers utility functions of users. In the context of cloud networking, when users utilize resources from a cloud provider, they incur a total cost to the cloud

TABLE III
A SUMMARY OF KEY FEATURES AND SUITABLE SCENARIOS OF AUCTIONS USED IN CLOUD NETWORKING.

Auction type	Market structure	Key descriptions	Suitable scenarios	Solution
English auction [93]	A seller, multiple buyers, and an auctioneer	Open-outcry ascending-price auction, and winning buyer pays the second highest price	<ul style="list-style-type: none"> Economics: seller's revenue maximization Cloud networking: bandwidth allocation 	Nash equilibrium
Dutch auction [93]	A seller, multiple buyers, and an auctioneer	Open-outcry descending price auction, and winning buyer pays the final price	<ul style="list-style-type: none"> Economics: the best price guarantee for buyer Cloud networking: bandwidth allocation 	Nash equilibrium
Vickrey auction/second-price sealed-bid auction [95]	A seller, multiple buyers, and an auctioneer	Sealed-bid auction, and winning buyer pays the second highest price	<ul style="list-style-type: none"> Economics: buyer's expected utility maximization Cloud networking: resource reservation, task allocation, and storage sharing 	Nash equilibrium
VCG [98]	A seller, multiple buyers, and an auctioneer	A generation of the Vickrey auction for multiple commodities, winning buyer pays a price equal to the loss of the social value due to its getting commodities	<ul style="list-style-type: none"> Economics: social welfare maximization Cloud networking: bandwidth allocation 	Bayesian Nash equilibrium
Double auction [99]	Multiple sellers, multiple buyers, and an auctioneer	Buyers and sellers submit respectively their bids and asks, and the auctioneer matches asks and bids	<ul style="list-style-type: none"> Economics: ordinary markets with multiple sellers and buyers which need to be cleared Cloud networking: bandwidth reservation, resource sharing, and task allocation 	Market equilibrium
Combinatorial auction [102]	A seller, multiple buyers, and an auctioneer	Buyers bid on combinations/packages of commodities, and winner determination problem is to find the optimal allocation of commodities.	<ul style="list-style-type: none"> Economics: markets where buyers compete on many different but related commodities Cloud networking: bandwidth allocation 	Optimal solution

provider. Therefore, the modified NUM problem which takes into account the total cost should be investigated. Consider the cloud resource reservation scenario in which N cloud tenants, e.g., video content providers, reserve network bandwidth from a cloud provider to deliver their videos to end-users. The goal of the cloud provider is to maximize the social welfare. Thus, the resource allocation problem can be expressed as

$$\begin{aligned} \max_{\mathbf{x}} \sum_i^N U_i(x_i) - C(\mathbf{x}) \\ \text{s.t. } x_i \in [a_i, b_i], i = 1, \dots, N, \end{aligned} \quad (8)$$

where x_i is the resource units that cloud tenant i receives, $\mathbf{x} = (x_1, \dots, x_N)$ is the vector of resource allocation, $U_i(x_i)$ is the monotonically increasing utility function associated with tenant i which is a strictly concave function of its resource allocation, $C(\mathbf{x})$ is a strictly convex function, and a_i and b_i are the constants.

Since the cloud provider has no knowledge of the utility functions, and the cost function is unknown to the cloud tenants, centralized methods such as interior point methods [120] which are typically applied to solve the NUM problems may not be applicable to the specific problem in (8). Alternatively, pricing-based iterative solutions are often used. Assume that the cloud provider charges cloud tenant i a price p_i for using resource x_i . Given a price vector $\mathbf{p} = (p_1, \dots, p_N)$, the problem in (8) is equivalent to

$$\max_{\mathbf{x} \in \prod_i [a_i, b_i]} \sum_i^N (U_i(x_i) - p_i x_i) + (\mathbf{p}^\top \mathbf{x} - C(\mathbf{x})), \quad (9)$$

where $(U_i(x_i) - p_i x_i)$ is the surplus of cloud tenant i for using x_i resource units, and $(\mathbf{p}^\top \mathbf{x} - C(\mathbf{x}))$ is the profit of the cloud provider. With the price vector \mathbf{p} , each cloud tenant selects x_i

resource units to maximize its surplus as follows:

$$x_i(p_i) = \arg \max_{x_i \in [a_i, b_i]} (U_i(x_i) - p_i x_i), i = 1, \dots, N. \quad (10)$$

Given the returned resource request x_i and applying the dual decomposition and gradient methods, the cloud provider updates iteratively the price vector \mathbf{p} according to the rule: $p_i = p_i - \gamma(y_i(\mathbf{p}) - x_i(p_i))$, where γ is an appropriate step size, and $y_i(\mathbf{p})$ is taken from $\mathbf{y}(\mathbf{p}) = (y_1(\mathbf{p}), \dots, y_N(\mathbf{p})) = \arg \max_{\mathbf{x} \in [a_i, b_i]} \mathbf{p}^\top \mathbf{x} - C(\mathbf{x})$. The process is repeated until the vector of resource allocation \mathbf{x} converges to the optimal resource allocation \mathbf{x}^* . Since (8) is a convex optimization problem, the solution \mathbf{x}^* is unique.

In practice, the above optimal solution is feasible only when utility functions of cloud tenants are concave. However, in delay-sensitive services, e.g., video and voice services, such utility functions vary with different types of services with inelastic flows. Therefore, the resource allocation may be a non-convex optimization problem [121].

IV. APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD DATA CENTER NETWORKING

Cloud networking provides local network connections to servers [122] (Fig. 1) and remote links for data center to create a resource pool supporting a large number of users and applications with diverse resource demands and utilities. Therefore, resource management becomes one of the most important issues in cloud networking. This section reviews applications of economic and pricing models for the resource management in cloud data center networking. The major issues include:

- *Bandwidth allocation*: Bandwidth allocation involves reserving and allocating bandwidth to users or application service providers. Traditional bandwidth allocation

algorithms often assume that the available resources do not change. However, in cloud networking, resources and demands can fluctuate randomly. Economic and pricing models have been used as the solutions in which all scarce resources can be best utilized with the variability of budget.

- *Request allocation*: Request allocation is to assign massive users' requests, e.g., the transaction and data processing, to data centers. Economic and pricing models provide efficient approaches to achieve the load balancing, latency and cost minimization taking network resource availability into account.
- *Workflow allocation*: The workflow allocation in the cloud data center networking is the computing task allocation among the data centers. Market-based approaches provide an efficient task assignment with the lowest cost and the fastest completion time through using negotiation mechanisms among network entities.

A. Bandwidth Allocation

As part of bandwidth allocation, the bandwidth reservation can be implemented to guarantee the bandwidth availability for users in future. This allows users to pay a lower price due to advance reservation. However, the bandwidth reservation may cause oversubscription or undersubscription issues [123]. This section reviews the applications of economic and pricing models for bandwidth reservation and allocation in cloud data center networking. Note that the bandwidth reservation is often implemented in advance compared with the bandwidth allocation which is done in an online basis.

1) *VCG auction*: A typical bandwidth reservation model in cloud networking is shown in Fig. 5. The model consists of a cloud provider, i.e., a seller, which owns a number of distributed data centers, and cloud tenants, i.e., buyers, which act as application and service providers. Cloud tenants rent bandwidth from the cloud provider to serve their subscribers. To avoid the high bandwidth reservation payment, the cloud tenants can lie about their revenues obtained by serving subscribers. Therefore, the authors in [124] adopted the VCG auction for the bandwidth reservation to achieve both optimal social welfare and strategy-proofness [125] that the cloud tenants have no incentive to lie about their revenue information. The cloud tenants simultaneously submit their bids to compete for bandwidth to the cloud provider. Each bid consists of bandwidth demands and the price per unit of bandwidth for which the buyer is willing to pay. To achieve the highest social welfare for the allocation, the winners are determined through a linear programming model which can be solved in polynomial time. The VCG mechanism was then applied to calculate the charge for each winner. The charge is the difference between the social welfare when the winner does not participate and when the winner participates in the auction. Since the proposed approach has an optimal allocation and calculates the charge based on the VCG auction, it was concluded to be a strategy-proof and optimal auction mechanism for cloud bandwidth reservation. The simulation results showed that both the social welfare and the bandwidth

satisfaction ratio of the proposed approach exceed 90% of the optimal solution when there are 200 tenants with 15 data centers.

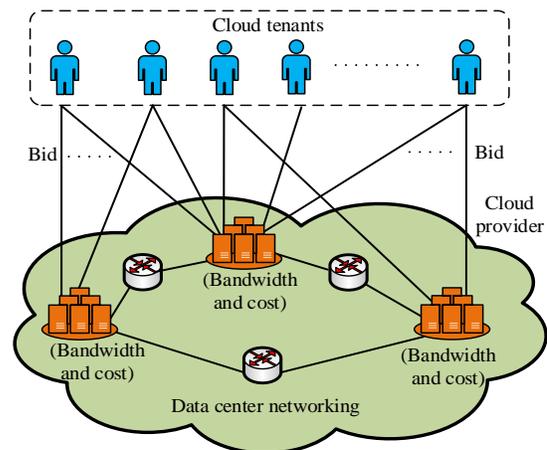


Fig. 5. A market based cloud bandwidth reservation in cloud data center networks.

2) *Shapley value based auction*: The VCG mechanism mentioned in [124] can simultaneously guarantee the truthfulness and the optimal economic efficiency. However, when the underlying allocation problem is NP-hard, e.g., the traffic scheduling, the VCG auction becomes computationally intractable. Approximation algorithms can be used, but they cause the VCG auction to lose its truthfulness [126]. To guarantee the truthfulness, approximation algorithms can be combined with suitable rules, e.g., by exploiting critical bids [127] or resorting to the linear programming decomposition technique [128]. In the context of bandwidth allocation for cloud users, the authors in [129] used the Shapley value method as the payment strategy instead of the VCG rule. In the auction, once receiving the data transfer requests (bids) from users, i.e., buyers, the cloud provider calculates a feasible traffic schedule to maximize social welfare based on linear programming. The Shapley value of each user is calculated as the average marginal charge by the cloud provider incurred by the user's traffic. Then, the cloud provider decides to reject or accept a user's request by comparing the user's bid price and its Shapley value. For example, if the user's bid price is larger than its Shapley value, the user's request is accepted, and its payment is equal to its Shapley value. Through the idea of cost sharing, the Shapley value approach was proved to be computationally efficient, budget balanced, individually rational, and truthful. However, the proposed approach only introduced the price. Other dimensions, e.g., path selection, need to be considered for each transfer task.

3) *Sealed-bid uniform price auction*: The above approaches only considered either bandwidth reservation [124] or bandwidth allocation [129]. The authors in [130] addressed both of them jointly through a two-tier pricing model with the aim of allocating efficiently bandwidth and maximizing the cloud provider's revenue. In the first tier, the reservation phase uses a premium price strategy to guarantee cloud tenants' minimum bandwidth ahead in time. A premium price, also

called image pricing, is a strategy of keeping the price of a product or service artificially high to encourage favorable perception among buyers [131]. The unallocated bandwidth remained from the reservation phase is traded in the second tier through a sealed-bid uniform price auction. The sealed-bid uniform price auction is a multiunit auction where a fixed number of identical units of a homogeneous commodity are set at the same price. The uniform price auction is applied due to its fairness in charging identical price for identical goods. In a single auction round in the second tier, the cloud provider formulates allocation functions based on the tenants' demand requests and bidding prices. Then, the cloud provider determines the market clearing price and allocates bandwidth based on these allocation functions. Finally, all winners pay the market clearing price for their respective allocated quantities. However, the proposed approach did not consider the future demands and the uncertainty in resource utilization of the reservation requests.

4) *Bargaining game*: The aforementioned auctions only satisfy either the buyer's or the seller's objective. To get a win-win solution for both, a cooperative game such as a bargaining game can be used. The authors in [132], [133] employed the bargaining game to address the rate allocation for VM-pairs in data centers as shown in Fig. 6. The VM-pairs, i.e., buyers, participate in an iterative bargaining game to negotiate the rates with the servers, i.e., sellers. Each VM-pair is associated with a utility gain which is assumed to be a convex function. The optimization problem is formulated to maximize the joint profit which is the product of buyers' utility gains. The problem is solved by the dual-based decomposition with the Lagrange multiplier method with the interpretation of rates and prices that buyers are willing to pay. The rate constraints are assumed to be linear [134]. Using the Karush-Kuhn-Tucker (KKT) conditions [135], the approach achieves a unique Nash bargaining solution of the rate allocation. The Nash bargaining solution ensures the Pareto optimality and achieves the fairness in resource allocation [136], [137]. The simulation results showed that the proposed approach can satisfy bandwidth demands of VM-pairs up to 99%. However, using the subgradient method results in slow convergence speed.

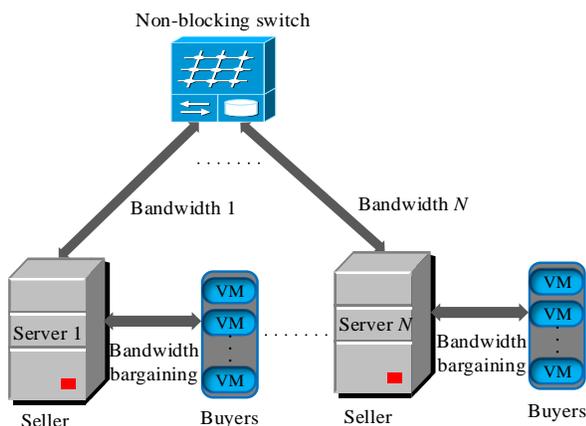


Fig. 6. A Nash bargaining game based rate allocation in cloud data center networks.

5) *Stackelberg game*: Auction approaches, e.g., sealed-bid auctions, allow the cloud provider to determine the price of bandwidth based on their bids. Typically, the cloud provider sets the bandwidth price to maximize its own revenue, and then the tenants decide their payments to obtain more bandwidth without paying too much. Thus, the interactions between the cloud providers and tenants can be modeled as a two-stage Stackelberg game as proposed in [138]. In the first stage, based on tenants' bandwidth demands, the cloud providers, i.e., leaders, cooperate with each other in a Nash bargaining game to optimize their pricing strategies. Different from [132], [134] which employed the subgradient method, found to have slow convergence to the optimal price, the authors in [138] proposed to combine the advantage of the demand segmentation method and the geometrical Nash bargaining solution [139]. Given the price and the amount of bandwidth that the cloud providers are willing to allocate, each selfish tenant, i.e., a follower, optimizes its bandwidth reservation through the weighted fair and max-min fair bandwidth reservation algorithms. The simulation results indicated that the proposed algorithms ensure the high average revenue of cloud providers and guarantee the bandwidth requirement for tenants.

In practice, the tenants' VMs of the same applications in a data center can be partitioned into the same virtual networks through network slicing [140]. The authors in [33], [141] adopted the Stackelberg game for allocating bandwidth to the tenants' virtual networks from the cloud provider. The cloud provider acts as the leader which announces a rental price per unit of bandwidth. Unlike [138], the leader's pricing strategy attempts to drive tenants' virtual networks, i.e., followers, to the social optimal solution and to make the highest profit for itself simultaneously. Due to the non-cooperative nature in terms of demand for network resources, virtual networks can be considered to be players in a non-cooperative game in which their strategies are to choose bandwidth demands to maximize their own utilities. Assume that the utilities are strictly concave functions of bandwidth and additive, the optimal bandwidth allocation among followers can be formulated as an optimization problem with multiple constraints. Through the distributed implementation, it was proved that the problem has a unique optimal solution which is the Nash equilibrium of the game. Given the set of optimal strategies of the followers, the leader then defines an objective function to be its own profit. Optimizing this function using the first-order conditions, there exists a unique Stackelberg equilibrium at which the leader's profit is maximized, and the bandwidth allocation for the followers reaches the Nash equilibrium [33]. Theoretically, the outcome of the Stackelberg game can achieve the better utilities compared with Nash equilibrium strategies which are obtained by assuming a simultaneous choice for leaders and the follower [142]. However, the experiment results to demonstrate this important comparison were not given.

6) *Differential pricing*: As stated earlier, the differential pricing applies different prices of the same product to different buyers. The authors in [59], [143] adopted the differential pricing for the allocation of bandwidth to tenants with the aim of maximizing the cloud provider's revenue. Specifically, the tenants, i.e., buyers, submit their requests to the cloud

provider, i.e., the seller. Each request has a bandwidth demand and a deviation factor. The deviation factor represents the flexibility in bandwidth allocation available to the provider. If two requests have different deviation factors, the tenant with a higher value gives higher flexibility to the provider than that with a lower value. The cloud provider gives more discounts, i.e., a lower price, to the tenant with the higher deviation factor. This flexibility introduces more degree of freedom of bandwidth allocation, leading to a higher efficiency. Simulation results showed that compared with the deterministic bandwidth allocation algorithm from [144], the proposed approach increases the allocated bandwidth up to 28% and the provider's revenue up to 12%. However, as shown in [145], the difference of the provider's revenue between the two approaches is slight when more discounts are given.

7) *Smart data pricing*: Smart data pricing adapts the resource price according to network congestion. This pricing strategy is used to create proper economic incentives for users. The authors in [146] adopted this strategy for allocating communication bandwidth to tenants' VMs when the VMs communicate with each other via multiple links as shown in Fig. 1. The objectives are to achieve the min-guarantee, i.e., guaranteeing the minimum bandwidth that the tenants expect for each VM, high utilization, i.e., maximizing network utilization, and network proportionality, i.e., the fairness among tenants. The experimental results showed that the proposed approach can satisfy up to 98% of tenants' VM demand compared with 76% of the baseline pricing from [147] which sets price based on only bandwidth. However, the proposed approach does not explain how to set the unit price of the min-guarantee bandwidth optimally.

The schemes to calculate the unit price of the min-guarantee bandwidth can be found in [148], [149]. The price is proportional to the number of serving VMs and is inversely proportional to the total sharing bandwidth. However, this pricing strategy is simple and does not ensure that the price is optimal.

8) *Other pricing models*: Apart from the common schemes described above, others pricing models have also been applied to the resource allocation in cloud networking.

Dominant resource pricing: Typically, cloud tenants rent both VM and network resource, i.e., a data transfer service. Therefore, the price that a tenant pays the data center owner, i.e., a cloud provider, depends on the time of VM occupancy and the size of the data transferred among VMs. In particular, the time of the VM occupancy depends on the location of the tenant because if the tenant is far from VMs' locations, the transfer time is longer. The authors in [150], [151], [152] proposed dominant resource pricing that applies prices independently of the tenants' locations, and thus reducing the tenants' costs. Since only the VM occupancy price depends on the tenant's location, the key technique is to enable the network price to dominate the VM occupancy price. Since the network price can be larger or smaller than the VM occupancy price, a bandwidth baseband threshold was introduced to guarantee that the bandwidth requirement is always larger than the base bandwidth. Therefore, the network price always dominates the VM occupancy price, and the tenants' payments are considered

to be flat, i.e., a location independent price. The simulation results showed that compared with the baseline pricing which sets price based on the task completion time, the tenants pay 70-80% less with the proposed pricing. However, the proposed approach does not specify how to set the bandwidth baseband threshold optimally.

Ramsey pricing: The authors in [66] exploited the Ramsey pricing [153] to regulate the demand of cloud tenants' bandwidth requests, achieving higher network utilization. The model consists of a cloud provider which owns a number of geographically distributed data centers connected with each other through a high-capacity network. Each virtual link between any two data centers is divided into several virtual pipes corresponding to several service classes. Each service class has a specific expected latency and a corresponding price which is set based on the Ramsey problem. The Ramsey problem pricing is a strategy in which a monopolist sets the price to maximize social welfare subject to the constraint on profit [154]. Applying this strategy to the proposed model, the price is set to maximize the welfare of both the cloud tenants and the cloud provider subject to the preset profit threshold. Solving this problem requires to have knowledge of the profit threshold and the expected demand for each service class. The threshold can be defined according to the market competition, and the expected demand can be obtained by cloud resource monitoring methods, e.g., a semi-centralized monitoring mechanism [155]. Generally, increasing the price leads to a decrease in the demand of that service class, or it will move the demand of that service class to other classes. The simulation results showed that the cloud provider's network utilization of the proposed approach is higher, e.g., 60% with the supply and demand pricing [156] or 40% with the static pricing. However, multiple cloud providers participating in the market need to be considered in the future work.

Pricing model based on dynamic programming: The above approaches did not consider the workload fluctuation of cloud tenants. Some events, e.g., releasing of a new movie, cause a high workload fluctuation. These applications require a continuous resource allocation elasticity to accurately adapt to the time-varying application's needs [157]. Therefore, it is important to predict the workload fluctuation. The authors in [158] proposed an approach to tackle this problem. The model consists of a cloud application owner, i.e., a buyer, which rents the inter-data center networking resources from a service provider, i.e., a seller. The application's requirements can be predicted by using a Markov chain model, utilizing the temporal variability of workload fluctuations [159]. The prediction allows the service provider to dynamically re-size the inter-data centers bandwidth pool to ensure the availability of network resources. Given the traffic workload fluctuation information, the service provider sets the resource prices at each time slot to maximize its long-term expected revenue. A dynamic programming algorithm [160] was adopted to search for the optimal prices. The simulation results showed that the mean square error between the estimated and the actual fluctuations can be marginal. Moreover, when the elasticity level, i.e., the bandwidth adaptation, is higher, the service provider receives a higher accumulated revenue since more

bandwidth demands are met at a higher price. However, more advanced techniques should be applied to accurately forecast workload fluctuations.

B. Request Allocation

The term “request” in this section represents the resource/workload/data processing requests from cloud users or tenants. A simple technique can be implemented by assigning each request of a user to the closest data center [161]. However, such a method can overload the data center during peak time and further degrade application performance as well as the revenue of the provider. Therefore, economic and pricing models have been developed to optimize the benefits of users and providers.

1) *Cost-based pricing*: To maximize the service provider’s profit while satisfying Service-Level Agreements (SLAs) of the users, the authors in [162] considered charging incoming user requests through the cost-based pricing. Cost-based pricing is a strategy to set the price of a request based on the cost of executing the request. In particular, whenever a new request from a user arrives at the cloud site, the service provider, i.e., the seller, estimates necessary parameters including network parameters (e.g., the bandwidth, control overhead, session throughput, and session lifetime), resource parameters (average inflow and processing speed), and data processing parameters (e.g., a job size, block size, and total time). These parameters allow the service provider to determine the SLA of the user and the resources allocated for the request. Then, the price of the request is defined according to the costs of resources. To maximize the service provider’s profit, the consumer perceive pricing is also employed through considering the money for which the user is willing to pay. However, other market factors, e.g., the market competition, need to be considered.

2) *Bargaining game*: To direct the users’ requests to the appropriate data centers, mapping nodes, e.g., authoritative DNS servers, can be deployed at different regions. The objective of this deployment is to optimize the general system performance. The mapping nodes cooperate for receiving the cloud resources from data centers. The authors in [85] modeled the data selection problem for users’ requests as the bargaining game among mapping nodes. Each mapping node, i.e., a player, has its own average utility as the objective which can be a convex function of the latency or throughput. Each node also has an initial utility which represents the minimum utility requirement for the users based on the SLAs with the cloud provider. The goal is to maximize the average utility cooperatively. Similar to [132], the optimization problem is formulated by maximizing the joint utility which is the product of nodes’ average utilities. However, unlike [132], to solve the non-linear problem, the sequential-linear-approximation algorithm [163] was employed for reducing computational complexity. The linear problem is then solved by the dual-based decomposition with the Lagrange multiplier method with the interpretation of load balancing, capacity, and cost constraints. Using the subgradient method, the load balancing and capacity constraints from data centers serve as price signals. When the total traffic routed to a data center exceeds its

capacity, the data center increases its price for the next round to suppress the excessive demand. This process continues until the algorithm converges to the optimal resource allocation. However, the slow convergence from using the subgradient method can impact the real-time applications.

3) *Non-cooperative game*: In a competition market with multiple service providers, a non-cooperative game can be used. The authors in [164] adopted the game for the request allocation among service providers with the aim of maximizing the social welfare of the system. The market consists of service providers, i.e., sellers, and a user, i.e., a buyer. Each service provider has its own control strategy which involves deciding on the number of servers placed in each data center and routing a user’s request to an appropriate server. The objective of each service provider is to minimize its operational costs while satisfying the users’ SLA and data center capacity constraints. Assume that each service provider’s strategy is kept private from other service providers. The set of optimal strategies yields a unique Nash equilibrium in which no service provider can optimize its cost by unilaterally changing its allocation strategy over time. The price of anarchy and the price of stability are then determined as the metrics to measure the best-case and worst-case efficiency loss of the game, respectively.

To tackle the worst-case where a service provider behaves selfishly in an uncoordinated manner, the authors in [165] extended the market with the participation of the cloud provider by taking into account a penalty function. The cloud provider must pay a penalty cost for the service provider when the cloud provider rejects the resource request from the service providers. This is reasonable since the request rejection can hurt the service satisfaction of users, resulting in the loss in service provider’s revenue. The experimental results showed that the proposed approach can improve the social welfare from 10% to 20% compared with the outcome of the competition game in [164].

C. Workflow Allocation

Workflow allocation in cloud network is to map required tasks from users to the resources at multiple network locations and order their executions so that task-precedence requirements are satisfied [166]. Due to the large resource distribution in the cloud network, the requirements for the workload allocation include minimizing the cost and delay of the task execution as well as adapting dynamically to the resource demand fluctuation from users. The traditional approaches exploited the diversity in local electricity prices, e.g., [167], [168], [169]. However, they focused only on minimizing the cost, i.e., Operational expenditures (Opex), of the operators rather than satisfying the users. Besides, the static approaches, e.g., [170], [171], may not achieve the objectives due to the lack of interactions among the entities in the cloud network. Therefore, the market based schemes with negotiation mechanisms among network entities have been developed to optimize the workflow allocation, resource utilization, and profit of the operators.

1) *Profit maximization*: The authors in [172] proposed a task scheduling algorithm across data centers through the economic model based on profit maximization as described in Section III-A3. Once receiving workload requests from users including the types of VMs and the number of time slots, the cloud provider solves the profit maximization problem. The total revenue depends on the number of tasks and their prices, and the costs are the Opex in the data center. The solution of the optimization problem allows the cloud provider to choose (i) an appropriate price for each type of tasks at each data center, (ii) the best number of servers to provision each type of VMs in each data center, and (iii) the optimal number of tasks of each type to schedule and to drop. Moreover, to enable the cloud provider to achieve a high time-average overall profit, the authors adopted the drift-plus-penalty framework in Lyapunov optimization [173]. It is a classical approach for translating a long-term time-average optimization problem into a series of similar one-shot optimization problems. The simulation results showed that the proposed pricing algorithm outperforms the static pricing, e.g., the pricing strategy in Amazon EC2's on-demand instance market, in terms of the profit. Moreover, the proposed approach achieves the stable profit over time.

2) *Spot instance pricing*: The approach in [172] aims at maximizing the cloud provider's profit. To reduce costs for users and achieve high utilization, the spot instance pricing can be used. Spot instance pricing which is practically used to set the price of the Amazon EC2 instances (<https://aws.amazon.com/ec2/spot/pricing/>) allows users to bid for idle or unused cloud instances. The users receive the instances immediately as long as their bids are higher than the spot price (i.e., the price of the spot instance) [174], which is often a discounted price. Different from the auction approaches, e.g., the first-price sealed-bid auction, the user only pays the spot price instead of its bidding price. The authors in [175] addressed the workload scheduling in cloud networks by combining the spot instance pricing and the on-demand instance pricing [176] to reduce the cost of workload execution while guaranteeing the workflow deadline. The model is shown in Fig. 7. Each user submits its bid to the task scheduler including the application task, current spot price, on-demand price, the current time, and factors that specify the weights of the types of pricing. The scheduler calculates the bid values and employs the checkpointing technique [177] which is known as an efficient fault tolerant strategy to find a suitable cloud resource for every task. If the spot instance is not enough for accomplishing the task, the scheduler adaptively switches to on-demand instances to meet workflow deadline. Experimental results showed that the proposed approach can reduce up to 70% execution cost compared with the task scheduling algorithm using only the on-demand instance pricing.

D. Applications of economic and pricing models for resource management in federated cloud networking

To guarantee SLAs for users' composite services, i.e., multi-cloud applications, cloud and network resources need to be distributed across multiple cloud providers [178]. Federated

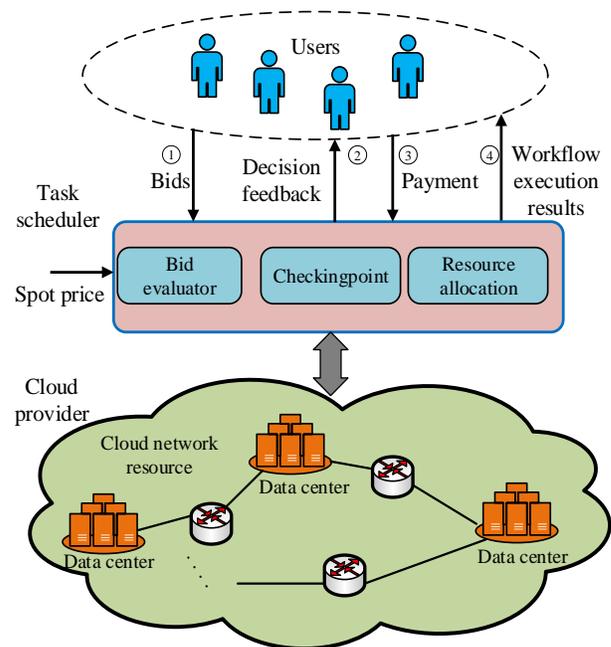


Fig. 7. Spot instance pricing based workflow allocation mechanism.

cloud networking (Fig. 8) enables interconnecting the cloud providers to form cloud federation resources which can be shared to increase capacity, availability, and resilience at multiple network locations [179]. The federated cloud network can be considered to be an extension of the cloud data center network belonging to the providers. Mathematically, it was also proved in [180] that the Return On Investment (ROI), i.e., the earned amount of money for each unit of investment, of a cloud provider in the federated cloud networking is higher than that in the stand-alone cloud data center network. In what follows, we discuss economic and pricing models which provide incentives to cloud providers to pool their resources in the federated cloud network to satisfy users' requests.

1) *Profit maximization*: To stimulate cloud providers within the federated cloud network for sharing resources, their profits need to be maximized. Therefore, the authors in [63] adopted the economic model based on profit maximization for allocating resource requests from users to cloud providers in a federated cloud network. The model consists of cloud providers interconnected through high capacity links. As shown in Fig. 8, each cloud provider receives resource requests from its local users and other providers. A cloud provider can insource, i.e., rent out, its unused resources to serve other providers, but it can also outsource, i.e., borrow, resources from other providers. To maximize cloud providers' profits and guarantee load balancing among them, each cloud provider determines the insourcing price as well as the available quota of resources. Using pricing policy proposed in [181], the insourcing price is set according to the VM costs, the fixed price to local users, and the maximum hosting capacity and idle capacity. Given the insourcing price, outsourcing prices (set by other providers), and networking costs (from the network providers), each provider solves the integer linear programming problem with the aim of maximizing the total profit of the providers

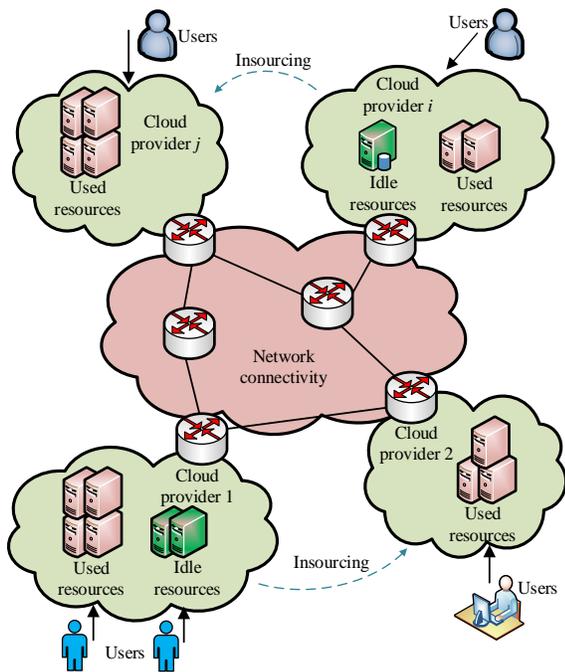


Fig. 8. Federated cloud network.

in the federated cloud network. The solution of this problem helps each provider to partition received requests into subsets that will be hosted locally or outsourced to other providers as well as to select the insourcing requests from other providers to accept. The simulation results showed that compared with the case without federation, the proposed approach can improve the profits of some providers up to 42% and the request acceptance rate up to 48%.

The approach in [63] did not specify how to set the resource prices. The authors in [182] determined upper and lower bounds for resource prices for both cloud providers and users. The model is shown in Fig. 9. First, each cloud provider calculates costs, e.g., energy costs, networking costs, deployment costs, and outsourcing costs. The lower bound for the insourcing price charging to the other providers is determined based on the condition to ensure that the revenue is higher than the overall cost. The upper bound for the insourcing price is obtained based on the proposed insourcing prices by the other providers, and particularly the upper bound is the minimum insourcing price among them. This enables one cloud provider to compete with the other cloud providers. Otherwise, the proposed resource price to users needs to be higher than the insourcing price to improve revenues of the providers within the federated cloud network. This condition allows to determine the lower and upper bounds for the proposed price to the users. These lower and upper bound prices are used as inputs for the revenue maximization problem of each cloud provider. The branch and bound algorithm [183] is used to solve the integer problem to find the optimal resource amount for serving the users, the amount for outsourcing and for insourcing. Nonetheless, the low complexity algorithm is needed.

The same approach can be found in [184] which determines

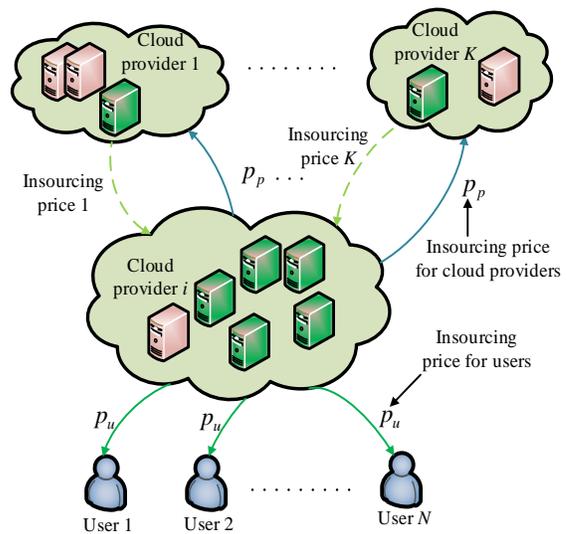


Fig. 9. Differential pricing based request allocation for federated cloud network.

a portion of job requests transferred from one provider to the other provider so that their total profit is maximized while guaranteeing the user utility. The model consists of two cloud service providers and their servers interconnected with each other through Internet links. First, each provider defines a profit function of its revenue and energy consumption cost. The revenue is based on the total incoming requests at the queue of the provider and the price per request. The total request of a provider includes the requests from its local users and the other provider's users. The price per request is set based on a QoS-dependent pricing scheme which should be convex and decreasing in the average delay of job execution. This delay is calculated according to the $M/M/1$ queueing model [185], taking into account the Internet transfer delay. The total profit maximization of the two providers is a non-linear optimization problem that is then solved with the method of Lagrangian and the KKT conditions. The solution is the optimal portions of the requests at each provider to be routed to the other provider [135]. The simulation results indicated that the proposed approach improves the total profit up to 100% compared with the case when each provider serves only its own users. However, the case with more than two providers that will make the problem more general was not considered.

2) *Differential pricing*: The aforementioned approaches did not provide any prediction techniques which address the workload elasticity of users. The future resource consumption of current users in the federated cloud network can be estimated through their behavior and probability of using the resource as proposed in [186], [187]. The resources considered here include virtual servers, computing and storage resources, virtual networks, and network bandwidth. The prediction can be implemented by a cloud broker which acts as an intermediary between the cloud providers and users. Basically, the resource estimation for the current user is based on the current resource price, the average of service relinquishing probability, and the history of relinquishing probability of the user. Relinquishing

probability of a user is the probability of giving up the resource for which the user has requested. Given the relinquishing probability, the differential pricing scheme was introduced to set different resource prices depending on the behavior of users [187]. Specifically, the proposed resource price is proportional to the service relinquishing probability of the user. For example, if the user has the high probability, the proposed price is set at a high value. This policy discourages the users with a high relinquishing probability from participating in the market because they may degrade the profit of providers. However, the proposed approach did not explain how to obtain the relinquishing probability for each user.

The same approach can be found in [188]. However, the authors in [188] investigated refunding the remaining amount to a user when its service consumption will be discontinued. The authors also considered that fog computing which is located between underlying Internet of Things (IoT) devices and the cloud computing can implement this task. The total refund is the sum of the refund of unutilized resources and the refund to be paid on quality degradation, i.e., not satisfying SLA. The refund is determined based on the amount of unutilized resources and a depreciation index. A high depreciation index allows the users which have used more services, e.g., more than 60%, to receive more refunds. The refund on quality degradation is based on the ratio of the acquired quality of service and the provided QoS.

3) *Cost-based pricing*: In the federated cloud network, users have to decide where to place their services on the federated cloud such that the service cost is minimized. The authors in [189] addressed the service placement optimization based on the cost model in federated hybrid clouds to guarantee the cost minimization for cloud users. A federated hybrid cloud (Fig. 10) is a composition of interoperable private and public cloud networks. Given a set of cloud services from the cloud providers, the total cost for each possible service placement option is evaluated by calculating the sum of the fixed costs and the variable costs. The fixed costs include the costs for hardware, e.g., servers and network devices, and software licenses. The variable costs comprise, e.g., the electricity cost, Internet connectivity cost, and cost of data transfer between different clouds. Finally, the cloud user selects the best option which has the minimum total cost. However, more performance factors, e.g., a service latency, need to be considered to guarantee user SLAs.

4) *Double auction*: Most of the above approaches consider the request allocation. The authors in [190] addressed the bandwidth reservation in federated cloud networking by using the double auction. The model consists of multiple cloud providers, i.e., sellers, multiple cloud tenants, i.e., buyers, and an auctioneer. The double auction used in this work is similar to that presented in Section III-B4. However, there are some slight modifications in the winner determination and payment stages. Specifically, upon receiving asks from sellers and bids from buyers, the auctioneer sorts sellers in the first list by the selling prices in a non-decreasing order. Then, the auctioneer sorts buyers in the second list by their bids in a non-increasing order. The water filling method [191] is used to fill the sellers one by one following the order in the first list, with buyers by

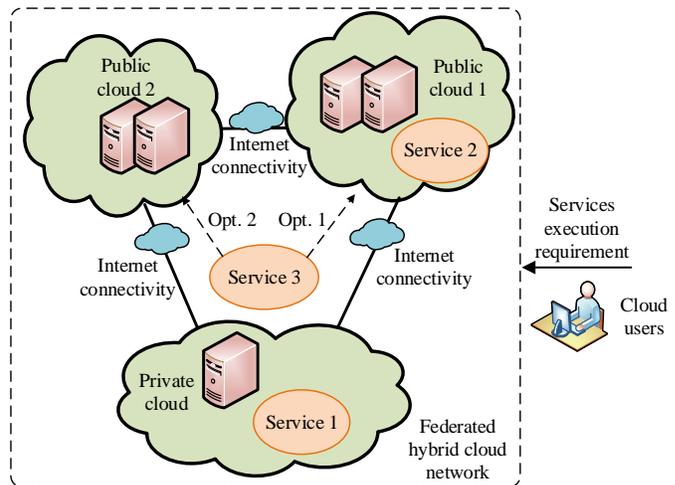


Fig. 10. Service placement in hybrid federated cloud network.

the order in the second list. The auctioneer finds the largest indexes k and l in the two lists to satisfy two constraints: (i) the first k sellers in the first list have sufficient bandwidth to satisfy the demands of the first l buyers in the second list, and (ii) the total charge to l buyers is less than total payment to k sellers. The second constraint is to guarantee the ex-post budget balance of the auction, i.e., the profit of the auctioneer is non-negative. The winners are the first k sellers and the first l buyers. The auctioneer pays each winning seller by the $(k+1)$ th selling price in the first list and charges each winning buyer by the $(l+1)$ th bidding price in the second list. Such payment and charging mechanisms are similar to those of the Vickrey auction that guarantee the truthfulness of the auction. However, the cheating behaviors of the bidders such as false-name bidding and collusion need to be considered in the future work.

Summary: In this section, we have discussed three major issues in cloud data center networks, i.e., the bandwidth allocation, request allocation, and workload allocation. We have then reviewed the applications of economic and pricing models for these issues which are summarized along with references in Table IV and Table V. We observe that in the cloud data center networking, economic and pricing models for bandwidth allocation get more attentions than other issues. In the federated cloud networking, the models aim at improving the total profit for cloud providers. In the next section, we review economic and pricing models in mobile cloud networking for efficient management of cloud and radio resources.

V. APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN MOBILE CLOUD NETWORKING

As stated in Section II, the Mobile Cloud Networking (MCN) is structured by several segments including cloud computing infrastructures (i.e., data centers), Radio Access Network (RAN), cloud mobile core network, and mobile platform services [26]. Since the wireless environment in MCN is dynamic, distributed, and heterogeneous, traditional static methods are not possible to adapt to achieve optimal

TABLE IV
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD DATA CENTER NETWORKING

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Bandwidth allocation	[124]	VCG auction	Cloud provider	Cloud tenants	Bandwidth	Based on buyers' bids, the seller determines winners by solving a linear program in polynomial time and charges each winner with the VCG payment	Resource efficiency, and welfare social maximization	Nash equilibrium
	[129]	Shapley value based auction	Cloud provider	Cloud users	Bandwidth	Based on buyers' bids, the seller compares each buyer's price with its own Shapley value to accept or reject the buyer. The accepted buyers pay their Shapley values for the seller	Social welfare maximization, computational efficient, individually rational, and truthful	Nash equilibrium
	[130]	Premium price and sealed-bid uniform price auction	Cloud provider	Users	Bandwidth	Buyers reserve bandwidth based on the premium pricing. The unallocated bandwidth remaining is traded using the sealed-bid uniform price auction	Maximum social welfare, and revenue maximization	Nash equilibrium
	[132] [134] [133]	Bargaining game	Servers	VM-pairs	Transmission rate	The unique optimal solution for the rate and price is determined by using the method of Lagrange multipliers and the karush-kuhn-tucker conditions	Min-bandwidth guarantee, high utilization, and fair allocation	Nash bargaining solution
	[138]	Stackelberg game	Cloud provider	Application provider	Bandwidth	Sellers set the bandwidth price based on the geometrical Nash bargaining game. The buyer optimizes its bandwidth reservation through the weighted fair and max-min fair reservation algorithms	High revenue, and bandwidth guarantee	Stackelberg equilibrium solution
	[33] [141]	Stackelberg game	Cloud provider	Tenants' virtual networks	Bandwidth	Seller uses the first-order conditions to set price, and buyers compete the bandwidth in a non-cooperative game	Efficient and fair allocation, and seller's profit maximization	Stackelberg equilibrium solution
	[59] [143]	Differential pricing	Cloud provider	Tenants	Bandwidth	Seller sets lower prices for buyers which accept to receive bandwidth allocation with high flexibility	Payment minimization, and high utilization	Pareto efficiency
	[146]	Smart data pricing	Cloud provider	Tenants' VMs	Bandwidth	Seller sets price of links between VMs depending on the congestion degree of the links to regulate VMs' demands	Min-bandwidth guarantee, high utilization, and network proportionality	Market equilibrium
	[148] [149]	Smart data pricing	Cloud provider	Tenants' VMs	Bandwidth	Same as [146], but the seller sets price of bandwidth reservation according to the number of serving VMs, and the total sharing bandwidth	Min-bandwidth guarantee, high utilization	Market equilibrium
	[150] [151]	Dominant resource pricing	Cloud provider	Cloud tenants	Data transfer service	Seller introduces a bandwidth threshold which the buyers must buy so that the network price dominates the VM occupancy price	Min-bandwidth guarantee, buyers' cost reduction	Flat-rate equilibrium
	[66]	Ramsey pricing	Cloud provider	Cloud tenants	Bandwidth	Seller sets different prices for different virtual pipes based on Ramsey problem	High utilization, social welfare maximization	Second-best equilibrium
	[158]	Dynamic programming based pricing	Service provider	Application provider	Bandwidth	Seller predicts the buyer's needs using a Markov chain and sets price for the network resource based on a dynamic programming algorithm	High resource adaptation, and high revenue	Optimal solution
Workload allocation	[172]	Profit maximization	Cloud provider	Users	Workload	Once receiving workload requests from buyers, the seller solves the profit maximization through the drift-plus-penalty framework in the Lyapunov optimization	Time-averaged overall profit maximization, and optimal accepted tasks	Optimal solution
	[175]	Spot instance	Cloud provider	User	Workload	Seller employs the checkpointing technique for scheduling the buyer's workload. The buyer pays the seller the spot instance	Buyer's cost minimization	Spot market equilibrium

resource management [19]. Economic and pricing models have been recently employed to dynamically and efficiently manage resources, e.g., the bandwidth and energy, in MCN, which are reviewed in the following.

A. Bandwidth Allocation

In this section, we review dynamic pricing schemes using auction mechanisms to address bandwidth allocation issues in MCN.

1) *Combinatorial clock auction*: The authors in [192] studied the method to assign efficiently wireless bandwidth to cloud users in MCN. The cloud users typically reserve the bandwidth for both uplink and downlink transmission within a specific time period. Thus, the combinatorial clock auction which is described in Section III-B4 is applied. The model consists of a mobile cloud network owner which acts as a seller (auctioneer), and cloud users which act as buyers (bidders). The auctioneer owns a set of spectrum bands. Each band may

have several bandwidth channels. The channels are auctioned within a specified time period and location. Bidders submit their package bids to the auctioneer. Each bid is a combination of channels in different bands, locations, and time periods and the corresponding price for which the bidder is willing to pay. The winner determination problem was formulated to identify the bids as winning or losing. The objective is to maximize the sum of the accepted bidding prices. The constraints include spectrum availability and duplex spacing between the uplink and downlink spectrum channels, i.e., paired spectrum channels. In particular, the duplex spacing constraint is to avoid the interference between the uplink and downlink channels. The spacing can be the amount of unpaired spectrum allocated to the bidder. The winner determination problem is NP-hard, and the anytime search algorithm [193] can be used which leads to a feasible solution at any time for the bandwidth allocation. However, the joint allocation of bandwidth and other network resources, e.g., base stations and

TABLE V
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD DATA CENTER NETWORKING (CONT.)

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Request allocation	[162]	Cost-based pricing	Service provider	User	Request execution	Seller sets price of the request execution based on the total cost of resources	Seller's profit maximization	Pareto efficiency
	[85]	Bargaining game	Data centers	Mapping nodes	Request directing service	A bargaining game among buyers is solved by the dual decomposition with the Lagrange multipliers. The balancing and capacity act as price signals to regulate the buyers' needs	Social welfare maximization	Nash bargaining equilibrium
	[164]	Non-cooperative game	Cloud provider	Service providers	Request execution	Sellers participate in the non-cooperative game in which their strategies are to decide the number of servers placed in each data center and routing users' requests to appropriate servers	Social welfare maximization	Nash equilibrium
	[165]	Non-cooperative game	Cloud provider	Service providers	Request execution	Same as [164], but sellers will receive a penalty cost if the cloud provider rejects the buyers' resource demands	Social welfare maximization	Nash equilibrium
	[63]	Profit maximization	Cloud providers	Users	Resource requests	Each seller sets resource insourcing price to other sellers based on its maximum hosting capacity and idle capacity. Maximizing total profit of federation providers is solved via the integer linear problem	Profit maximization	Optimal solution
	[182]	Profit maximization	Cloud providers	Users	Resource requests	Each seller proposes upper and lower bounds for resource prices to other sellers and buyers. Then, the branch and bound algorithm is used to determine optimal resource allocation	Profit maximization, and optimal request allocation	Optimal solution
	[186] [187]	Differential pricing	Cloud providers	Users	Service requests	A broker considers behaviors of buyers through defining their relinquishing probabilities to set different prices	Efficient resource allocation, and profit improvement	Pareto efficiency
	[189]	Cost-based pricing	Cloud providers	Cloud user	Service requests	Buyer evaluates the total cost of all possible resource options for its service execution from different sellers through calculating fixed costs and variable costs. The option with minimum total cost is selected	Buyer's cost minimization	Pareto efficiency
	[190]	Double auction	Cloud providers	Cloud tenants	Bandwidth	The auctioneer uses the water filling method to select the winning buyers and winning sellers. The charging and payment policies are based on the Vickrey auction	Truthfulness, and ex-post budget balance	Market equilibrium

backhaul links, needs to be investigated.

2) *Conventional auction*: The auction mechanism as proposed in [192] requires a high communication overhead. Alternatively, the Dutch auction (see Section III-B4) can be used for the bandwidth redistribution as proposed in [194]. The model is shown in Fig. 11 in which there are one cloud service provider, i.e., a seller, and multiple interfacing gateways, i.e., buyers. At a time slot, each gateway has a number of mobile users connected with it with the QoS guarantee. Initially, the service provider sets the ceiling price per unit of bandwidth and broadcasts the price to all gateways. Each gateway submits a bid to the service provider including the minimum required bandwidth which guarantees the QoS, i.e., the service delay, for its mobile users. The service provider compares its maximum bandwidth availability and the total demand of the gateways. If the available bandwidth is greater than or equal to the total demand, the service provider terminates the auction and allocates the bandwidth proportionally to the demands of gateways. Otherwise, the service provider decreases the price and broadcasts it in the next time slot. The auction continues until the total demand is greater than or equal to available bandwidth. Each gateway then pays the service provider the price of bandwidth based on the prepay mechanism as in [195]. In general, a gateway tries to maximize its utility which is the difference between the revenue that it receives from serving mobile users and the price that the gateway pays to the service provider for the allocated bandwidth. The utility is a concave function with respect to the gateway's bid. Using the second-order derivative of the utility function, it was proved that there

exists a Nash equilibrium at which the gateways' utilities are maximized. However, the assumption that each gateway knows the bid value of others is not realistic.

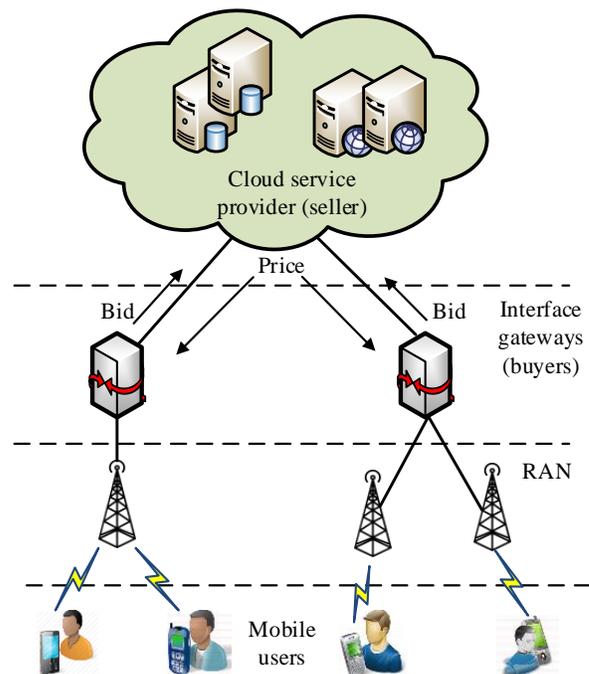


Fig. 11. Bandwidth allocation based on auction in MCN.

One shortcoming of the approach in [194] is that the service provider may redistribute bandwidth for all gateways even if

TABLE VI
A SUMMARY OF ADVANTAGES AND DISADVANTAGES OF MAJOR APPROACHES FOR THE RESOURCE MANAGEMENT IN CLOUD DATA CENTER NETWORKING.

Major approaches	Advantages	Disadvantages
[124]	<ul style="list-style-type: none"> Support multiple bandwidth units 	<ul style="list-style-type: none"> Have high computational complexity
[129]	<ul style="list-style-type: none"> Achieve computational efficiency 	<ul style="list-style-type: none"> Evaluate bids based on only a single attribute
[130]	<ul style="list-style-type: none"> Support both bandwidth reservation and bandwidth allocation 	<ul style="list-style-type: none"> Support a single bandwidth unit Have uncertainty in future demands and resource utilization
[132]	<ul style="list-style-type: none"> Achieve win-win solution 	<ul style="list-style-type: none"> Have slow convergence
[59]	<ul style="list-style-type: none"> Adapt to the high flexible requests of users 	<ul style="list-style-type: none"> Not adapt time-guarantee requests
[66]	<ul style="list-style-type: none"> Achieve high network utilization 	<ul style="list-style-type: none"> Support only a single cloud provider
[158]	<ul style="list-style-type: none"> Adapt to the time-varying application's needs 	<ul style="list-style-type: none"> Have high computational complexity
[172]	<ul style="list-style-type: none"> Achieve stable profit over time 	<ul style="list-style-type: none"> Have high computational complexity
[175]	<ul style="list-style-type: none"> Achieve low computational complexity Guarantee workflow deadline 	<ul style="list-style-type: none"> Have unreliability in spot instances
[162]	<ul style="list-style-type: none"> Be easy to be implemented 	<ul style="list-style-type: none"> Not use market factors
[85]	<ul style="list-style-type: none"> Achieve win-win solution 	<ul style="list-style-type: none"> Have slow convergence
[186]	<ul style="list-style-type: none"> Address the workload elasticity of users 	<ul style="list-style-type: none"> Be challenging to determine of users' behavior
[190]	<ul style="list-style-type: none"> Support multiple cloud providers and multiple cloud tenants 	<ul style="list-style-type: none"> Have unstable equilibrium Have slow convergence

only one gateway is overloaded by many users. The authors in [196] investigated sharing users among an overloaded gateway and its neighbors. The English auction is used. Initially, the overloaded gateway, i.e., the seller, broadcasts the information of the user's location to neighbors, i.e., buyers. If the user is within the region of a neighbor, this gateway sends the overloaded gateway an average QoS index which is determined based on its allocated bandwidth and the service delays of current users. The overloaded gateway also estimates a QoS index for the user based on the gateway's average QoS index and the information received from the service provider including the minimum delay threshold and the service delay. The overloaded gateway accepts the interested gateways as participants in the auction if the difference between its estimated index and the neighbors' average index is smaller than a threshold. At the initial state of the auction, the overloaded gateway broadcasts the minimum price for sharing the user. In each iteration, the gateway increases the price until its current utility value exceeds the initial value. The overloaded gateway allocates the user to the neighbor which accepts the price. The overloaded gateway then gets the payoff from the winning neighbor. However, the assumption about the information provided by the service provider for estimating the QoS index is not always possible.

B. Resource management in Cloud-RAN

Cloud-Radio Access Network (Cloud-RAN) is a centralized, cloud computing-based architecture for radio access networks [197]. In Cloud-RANs, signal processing functions of a base

station is performed in the cloud, i.e., centralized BaseBand processing Units (BBUs) or BBUs pool, as shown in Fig. 12. Then, the transmissions of radio signals to users are performed by Remote Radio Heads (RRHs) based on the baseband signals received from the cloud. To connect BBUs and RRHs, fronthaul links are used. One of the design goals in Cloud-RAN is to minimize the total downlink transmission power from RRHs to users while maintaining the fronthaul capacity and user QoS constraints.

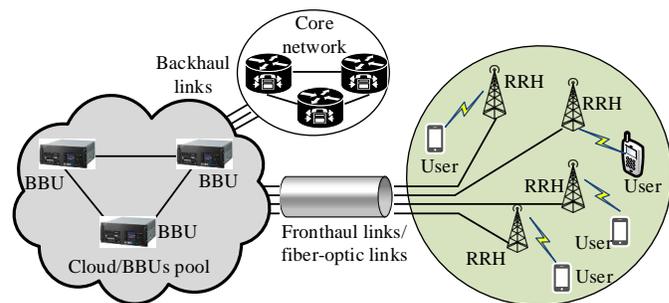


Fig. 12. Cloud-RAN architecture.

To achieve the goal, the authors in [198] studied two problems. The first problem aims to determine a set of RRHs serving each mobile user and the precoding vectors for RRHs to minimize the total transmission power from RRHs to mobile users with the constraints on fronthaul capacity. The second problem is to minimize the total transmission power from RRHs to mobile users and total fronthaul capacity between BBUs and RRHs. The first problem is addressed by iteratively

solving the second problem while adjusting pricing coefficients associated with RRHs. A pricing coefficient for each RRH refers to the price per unit of fronthaul capacity for the link between the cloud and RRH. The fronthaul capacity of the RRH is a decreasing function of its pricing coefficient. Solving the first problem is implemented via the binary searching method which adjusts the pricing coefficients so that the fronthaul capacities of RRHs are equal to their maximum allowable limits. In the second problem, the objective function is concave, and the feasible region corresponding to all the constraints is also convex. Thus, this problem can be solved by using the gradient method. Simulation results showed that the total transmission power is smaller when the maximum number of RRHs serving one user is larger. However, this also results in the high computational complexity.

The model in [198] consists of a single cloud serving multiple users via RRHs. Multiple clouds can be used to satisfy the processing demands of users, called M-CRAN (MultiCloud RAN) [199]. The authors in [200] addressed the issue of assigning users, i.e., buyers, to clouds, i.e., sellers, such that the overall net benefit of each cloud is maximized. Each cloud solves the knapsack problem [201] the objective and constraint of which are the net benefit and the resource budget, respectively. The resource budget is defined as the maximum number of users that the cloud can serve. The cloud also pays its users penalty costs if the QoS service cannot be guaranteed. Therefore, the net benefit function of the cloud serving a user is the difference between the price that the user pays and the penalty cost. The optimization problem is NP-hard. However, a full polynomial time approximation scheme [202] can be used to find an optimal set of users. The cloud can increase the penalty cost for attracting more users and iteratively perform the algorithm to maximize its overall net benefit. However, the high computational complexity will be incurred.

Summary: In this section, we have reviewed the existing literature of pricing-based resource management approaches in MCN. These approaches with their references are summarized in Table VII. As seen, auction-based approaches are well suited for the bandwidth allocation in MCN. However, the pricing models developed for the resource management in Cloud-RAN are relatively few. Further research is required to extend the preliminary results reviewed in this section. In the following section, we review the existing economic and pricing models for resource allocation in edge computing which includes a variety of network technologies, i.e., the cloudlet, cloudIoT, social cloud, and self-organization cloud.

VI. APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN EDGE COMPUTING

The cloud data center and mobile cloud networking are considered to be centralized paradigms, with storage and processing resources hosted within large data centers belonging to cloud providers. However, such paradigms face issues such as peak usage, high operational costs, bandwidth bottlenecks, and service interruption due to natural disasters (e.g., fire, earthquake, and power outage) [53]. Edge computing models

that exploit distributed “edge” devices can solve the issues. As shown in Fig. 13, edge devices can be small-scale data centers, volunteered computers, users devices (e.g., laptops, smartphones, and iPads) and sensors that are at the periphery of the network. Therefore, edge computing pushes the frontier of computing applications, data, and services away from the core of the data center network to the edges [38]. To attract users to contribute their resources, incentive mechanisms using pricing and payment strategies have been adopted in order to guarantee the stable scale of participants and QoS. Thus, this section reviews economic and pricing models for the resource management in some common models of edge computing.

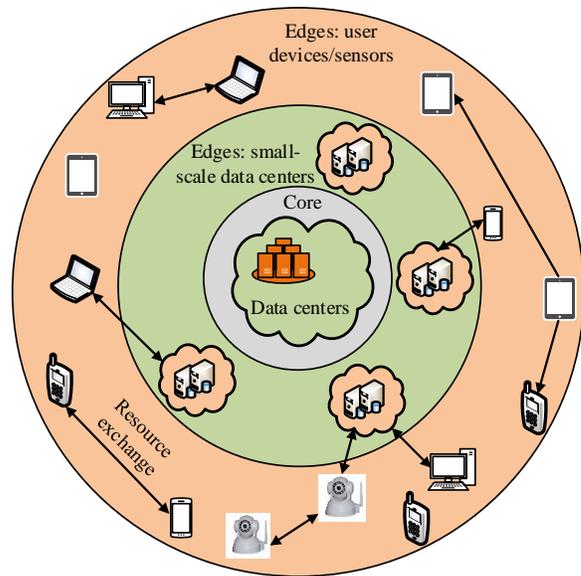


Fig. 13. Edge-centric computing model.

A. Cloudlet

Cloudlet, also known as mobile edge computing [203] or mobile micro-cloud [204], is a mobility-enhanced small-scale cloud data center. It can be located at the edge of the network, e.g., at the base station to which the mobile users connect, with the aim of providing cloud services to mobile users at lower latency [205]. Thus, the cloudlet is considered to be the middle tier of a 3-tier model, i.e., the mobile device-cloudlet-cloud as shown in Fig. 14. Resources at the cloudlet tier are limited [206]. Therefore, competition-based pricing models such as auction, non-cooperative game, or supply-demand model are effectively used for resource allocation to mobile users.

1) *Double auction:* To motivate resource pooling, cloudlets can form cloudlet groups as illustrated in Fig. 14. The authors in [207] adopted the real-time group-buying auction for the cloudlet group to offer its services, i.e., mobile videos, to nearby mobile users with lower prices while maximizing the profit of the group. The group-buying auction is a type of double auction in which buyers get more discounts from sellers if more buyers participate [208]. The model consists of the cloud (i.e., the supplier) connected to the cloudlet group (i.e., the retailer) through the Internet and the mobile users (i.e.,

TABLE VII
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN MOBILE CLOUD NETWORKING

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Resource management	[192]	Combinatorial clock auction	Mobile cloud network owner	Cloud users	Bandwidth	Based on buyers' bids, the winner determination problem is solved by the anytime search algorithm	Efficient allocation	Optimal solution
	[194]	Dutch auction	Cloud service provider	Gateways	Bandwidth	Based on buyers' bids, the seller compares its bandwidth availability and the total demand of buyers to either allocate resources or decrease the price such that the demand and the supply are equal	Buyers' utility maximization	Nash equilibrium
	[196]	English auction	Overloaded gateway	Neighboring gateways	Sharing user	Seller initially broadcasts the minimum price for sharing user to buyers, and then increases the price until its current utility exceeds the initial value. The buyer which accepts this price is selected as the winner for serving the sharing users	QoS guarantee, and seller's utility maximization	Nash equilibrium
	[198]	Generic pricing	Cloud/BBU	RRHs	Fronthaul capacity	Seller sets the price per unit of fronthaul capacity. The binary searching method is used to adjust the price to solve the problem which determines a set of RRHs serving each mobile user and the precoding vectors for RRHs	Buyers' power minimization, and optimal trade-off between transmission power and required fronthaul capacity	Optimal solution
	[199]	Knapsack problem	Clouds	Users	Service	Each seller solves the knapsack problem to select an optimal set of buyers which maximize its overall net benefit. The problem is solved by the full polynomial time approximation method	Seller's benefit maximization, and QoS guarantee	Pareto efficiency

TABLE VIII

A SUMMARY OF ADVANTAGES AND DISADVANTAGES OF MAJOR APPROACHES FOR THE RESOURCE MANAGEMENT IN MOBILE CLOUD NETWORKING.

Major approaches	Advantages	Disadvantages
[192]	<ul style="list-style-type: none"> Require little global information Achieve economic efficiency 	<ul style="list-style-type: none"> Have high computational complexity
[194]	<ul style="list-style-type: none"> Have fast convergence 	<ul style="list-style-type: none"> Require a centralized allocation algorithm
[199]	<ul style="list-style-type: none"> Support multiple cloud providers 	<ul style="list-style-type: none"> Have high computational complexity

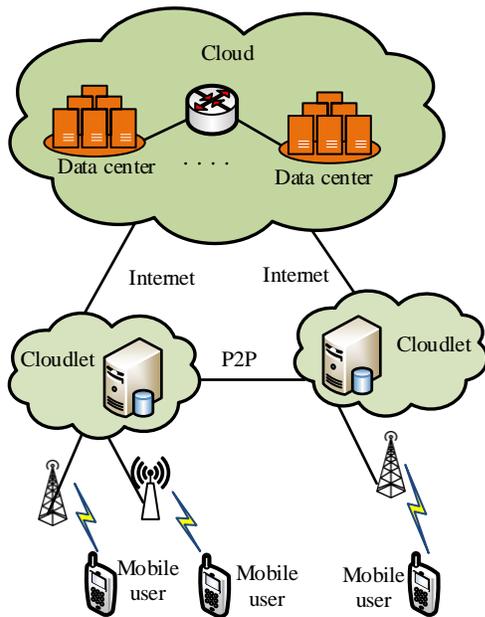


Fig. 14. A 3-tier model with mobile devices, cloudlets, and cloud.

period. Users with bidding prices higher than the auction price are selected as successful bidders, i.e., winners. Potential bidders including unsuccessful bidders and new incoming users are sorted in a descending order of bidding prices. The cloudlet then finds the first bidder with the bidding price not less than the auction price in the price curve. The price curve is a non-increasing sequence of auction prices obtained via maximizing the expected profit function of the cloudlet. Potential bidders, the bidding prices of which are higher than that of the first bidder, are the winners. When the auction ends, all winners buy the services at the final deal price, which is lower when more winners are selected. This strategy stimulates more users to choose the service from the cloudlet group.

The double auction discussed in [207] can achieve the individual rationality and budget balance, but it cannot guarantee the truthfulness. The authors in [209] addressed this problem by charging users according to the payment policy of the Vickrey auction. The model consists of: (i) mobile users, i.e., buyers, (ii) cloudlets, i.e., sellers, and (iii) a central controller, i.e., an auctioneer. The cloudlet serves only its nearby mobile users to reduce communication latency. The auctioneer sorts buyers in an ascending order of bids and sellers in a descending order of asks. The ask of the median seller is selected as a threshold to determine the winning buyer and seller candidates. For each winning seller candidate, the

buyers). One of the cloudlets starts the auction with an initial auction price, a specified supply quantity, and an auction

auctioneer selects a winning buyer with the highest price and charges it a price of the second highest bid. If the buyer wins two or more sellers, the auctioneer can select only one seller such that the buyer's utility is the highest. The simulation results showed that when a buyer bids a truthful price, its utility is improved. However, the system efficiency in terms of the number of final matchings between winning buyers and winning sellers only achieves around 50% of that of the optimal strategy.

Using the same model as in [209], the authors in [210] considered the randomness and the uncertainty in the auction to improve the system efficiency. Specifically, the auctioneer sorts sellers randomly as a list. To determine a winning buyer for each seller, the auctioneer defines the ask vector excluding the ask of that seller and then calculates the median ask of this vector. Among buyers with the bids higher than the ask of the seller, the buyer with the highest bid wins the service of the seller. Then, the winning buyer and the seller are inserted to the sets of winning buyers and winning sellers, respectively. The clearing price charging to the winning buyer and the price paid to the seller are set to be the same. More specifically, the price is the maximum of the median ask and the second highest bid of all buyers for the seller. Since any winning buyer in the set of the winners does not compete with other buyers for the remaining sellers, the candidate elimination algorithm as in [209] is not necessary, and the system efficiency is thus improved. The simulation results showed that the proposed solution achieves the system efficiency up to 80% of that of the optimal strategy. However, the proposed solution cannot guarantee strong truthfulness for buyers.

2) *Non-cooperative game*: The authors in [72] considered a model with multiple brokers which assign cloud resources, i.e., the computation resource and network bandwidth, reserved from the cloudlet and the public cloud to mobile users as shown in Fig. 15. Long-term reservation and on-demand request are applicable at the public cloud, but the bid proportion policy should be implemented at the cloudlet due to its limited resources. The bid proportion policy [211] allocates resources to buyers proportionally to their purchasing prices. Each broker, i.e., each buyer, decides its bidding price and on-demand request such that its average cloud price is minimized, given other brokers' strategies. The average price is a convex function of the bidding price, and the brokers are selfish to minimize the brokers' average prices. Therefore, the non-cooperative game is used to determine their optimal decisions. The Jacobi best-response algorithm [212] is then adopted to iteratively achieve an approximation of the Nash equilibrium at which bidding prices of all brokers are optimal. Simulation results showed that the proposed solution can reduce the price around 23% compared with the case that mobile users submit their requests to the public cloud directly. However, information sharing schemes among brokers to enforce the truthfulness need to be developed.

3) *Supply and demand model*: The above models are for static environments in which users are not on the move. When they are moving, pricing strategies relying on central entities, e.g., the auction, are not appropriate. In such a system, the authors in [213] addressed the issue of exchanging cloudlet

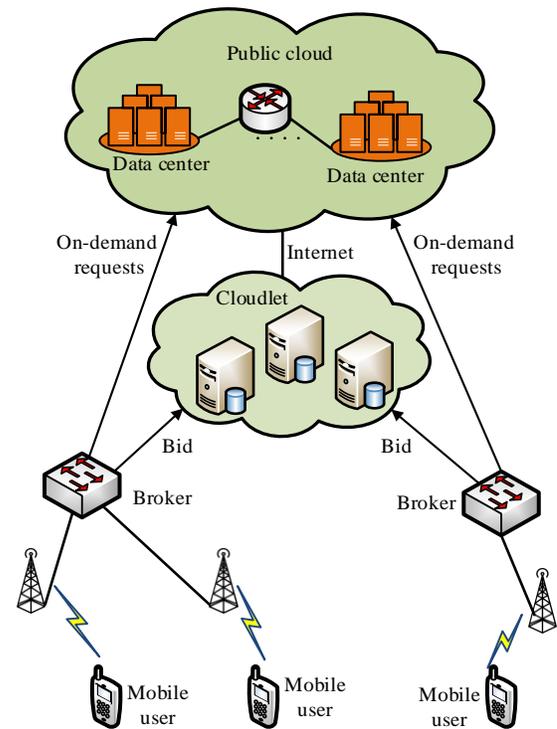


Fig. 15. Resource allocation in presence of brokers.

resources including CPU cycles, storage, and broadband network, among mobile users through the supply and demand model. Each user owns a cloudlet and acts as a buyer and a seller during its mobility. Given the budget, the objective of each user is to maximize the individual payoff which is the difference between the utility of buying resources and the cost of selling resources. This is a convex problem and can be solved by the primal-dual algorithm [214]. The solution allows the user to decide an optimal amount of resources to sell. To clear the resource market while maximizing each user's utility, an aggregate excess demand function is introduced. This function is the difference between the total demand and the total supply over all users in network. The classical tatonnement process [215] is used for the price adjustment to achieve the market equilibrium where the total demand equals the total supply. In particular, if the total demand exceeds the total supply, the seller increases the unit resource. Otherwise, the unit price should be reduced. The process continues until the function equals zero. However, the equilibrium may not be stable, and the computational cost for converging to the equilibrium can be high.

Apart from the aforementioned cloudlet models, a similar model, called Mobile Telecom Cloud (MTC), is found in [216]. Edge cloud services are given by mobile network operators which provide the last-mile Internet access to mobile users as shown Fig. 16. The network operators act as brokerages which use discount from cloud providers, e.g., Amazon, to offer better and cheaper cloud services to their users. Specifically, when receiving users' cloud requests, the brokerage formulates the resource reservation as the total cost minimization problem. The total cost depends on the users'

cloud requests, costs of cloud services offered by the brokerage or cloud providers, and the discount threshold from the cloud providers. This problem is then solved by either the linear programming combined with the rounding technique [217] or the min-cost greedy. The optimal solution allows the brokerage to set its price range using two conditions: (i) its offer price is less than the proposed price of the cloud provider, and (ii) the total cost of the brokerage is less than the sum of charged prices to users. These conditions aim to guarantee a high profit for the brokerage while attracting more users. The simulation results indicated that the cost of brokerage with the min-cost greedy algorithm is smaller than that obtained from with the linear programming. Moreover, the min-cost greedy algorithm runs much faster than the linear programming with rounding which requires multiple iterations to converge.

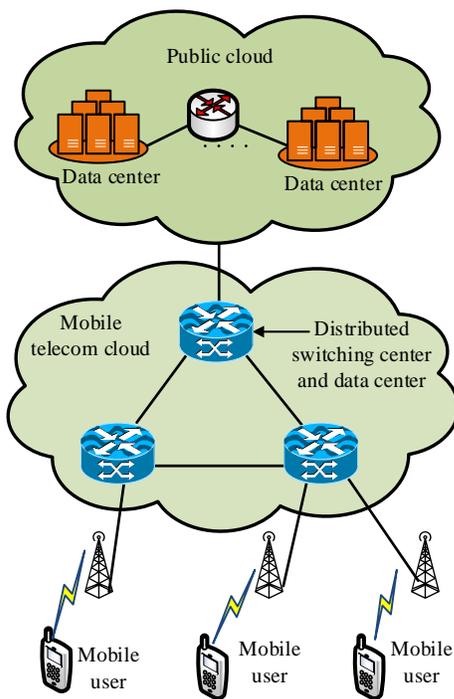


Fig. 16. Mobile telecom cloud model.

B. Volunteer Computing Systems

This section reviews economic approaches for resource allocation in volunteer computing systems. Similar to the cloudlet model, volunteer computing systems allow computer owners to contribute their computing resources for processing users' tasks. However, instead of using small-scale data centers as in the cloudlet model, a large number of distributed volunteered computers are used. They are connected with each other over WANs as illustrated in Fig. 17. Some projects are recently designed based on the framework, e.g., BOINC [218] with 65,000 computers, and Cloud@Home (<http://clouds.gforge.inria.fr/pmwiki.php>). In such a network, any volunteered computer, also called a host or a node, can act as a task scheduler and a resource contributor to schedule and execute users' tasks.

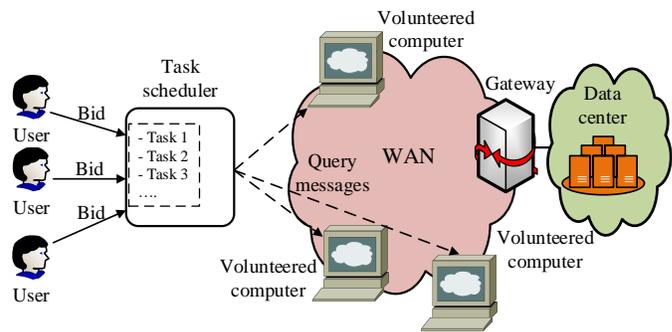


Fig. 17. Task scheduling in volunteer computing system.

1) *Reverse Vickrey auction*: Like the cloudlet model, due to the limited resource at the volunteered computers, the reverse Vickrey auction is usually applied for allocating users' tasks to the volunteered computers in volunteer computing systems. Such an approach was proposed in [219]. More specifically, when receiving a task request from a user, the task scheduler sends a query message including the expected qualified resource demand of the task, e.g., the CPU, memory, and network bandwidth, to its neighbors using the distributed range query protocol [220]. This searching continues until the number of hops is greater than a time-to-live threshold. Resource nodes (sellers) reply the task scheduler (buyer) the response messages (asks) containing the resource nodes' identifiers (e.g., IP address), resource availability states, and the corresponding prices. The resource node with the lowest price is selected for executing the user's task, and the payment for this node is implemented according to the Vickrey auction policy. Compared with the random-node-selection approach, the simulation results indicated that the proposed approach improves the resource node's payoff up to five times. However, the finished task ratio of the proposed approach is slightly lower than that of the random selection approach. A multi-attribute reverse auction can be used to improve the performance.

2) *Profit maximization*: Since the task scheduler acts as a broker to map task requirements to the resources of nodes, the profit of the broker needs to be considered. The authors in [221] employed the profit maximization-based pricing to determine the costs paid for the volunteered computer owners and the resource prices offered to the users to maximize the broker's profit. Each user submits to the scheduler a task execution request including the required resources (i.e., the computation and network resources), the number of time slots, the deadline, and a desired success probability of the reservation in the case of resource failures. The scheduler determines the probability distributions of price acceptance for users and the probability distributions of cost acceptance by the owners. These distributions can be learned in an online fashion. Based on these distributions, the broker defines its profit by computing the total revenue minus the total cost. The optimization problem is then solved by a sequential optimization algorithm in which each price and cost for a request type is optimized sequentially. Simulation results showed that the total profit of the broker at the low demand is significantly

higher than that at the high demand. The reason is that even at high demand, the broker cannot increase the price offered to users to ensure that they still accept the prices.

3) *Demand-based pricing*: The approaches in [219], [221] may not satisfy some users which are willing to pay proportionally to the QoS. Hence, the authors in [222] adopted the demand-based pricing for allocating users' tasks to resource nodes. The demand-based pricing charges users according to their demand. This pricing strategy is practically adopted by some cloud service providers, e.g., CloudTweaks (<http://cloudtweaks.com/>), to maximize the resource utilization and guarantee low costs. Specifically, given users' resource requirements and nodes' budgets for executing tasks, the node sets the resource price based on the total resource demand of tasks across the network. If the total demand is less than 50% of the total capacity of the node, the price is charged according to the base price. Otherwise, the higher price is applied. Also, the node uses the k -means clustering algorithm to classify the tasks into high, medium, and low priority levels. The node checks the resource requirement of the task with the highest priority. If the node has available resource to serve the task and the price for executing the task is within the budget constraint, the task is given with the preference for execution in the node. Compared with the first-come first-served approach, the proposed approach improved significantly the throughput ratio while reducing the average payment of users. However, this may lead to revenue loss of resource owners.

4) *VCG auction*: In practice, users require not only resources within an internal system, but also external resources, e.g., data center resources. They need to access the external resources via a gateway of a bandwidth provider [223]. The authors in [224] considered allocating bandwidth to users so as to maximize the provider's revenue. Additionally, it must ensure the social welfare maximization for the users even if they can lie about their priority to get higher utility. The VCG auction is used to achieve this goal. Users, i.e., buyers, submit to a provider, i.e., a seller, their bids, each of which specifies the priority class, the bandwidth demand, the valuation, and the price. Given these information, the provider defines each user's utility. The provider finds an optimal schedule to maximize the sum of utilities of all users through the use of the greedy approach. Each user is then charged according to the VCG payment policy. The simulation results highlighted that when varying the probability of user lying, the social welfare of the VCG auction is always significantly higher than that of the first-price and the Vickrey auctions.

C. Client-Assisted Cloud Systems

Client-assisted cloud models are distributed cloud paradigms which form resource pooling by exploiting resources of clients [38]. Here, a client or user is referred to as an "edge" device belonging to the external network environment of the edge-centric computing as shown in Fig. 13. Such paradigms aim to reduce the network traffic and resource burden at servers in volunteer computing systems, cloudlets, and data centers [225]. In particular, in what follows, we review economic approaches which have

been used to incentivize users/clients to contribute their local resources in the different distributed cloud models.

1) *Client-assisted cloud storage system*: The authors in [226] addressed the issue of constructing a storage pool for storage service providers, e.g., Amazon S3, using under-utilized storage and network bandwidth resources of cloud users as illustrated in Fig. 18. Due to asynchronous arrivals of users and service provider [227], the online reverse auction can be applied. Note that the auction is commonly used for the online e-commerce, e.g., eBay (<http://www.ebay.com/rpp/live-auctions>). Accordingly, users, i.e., sellers, submit their asks to the service provider, i.e., the buyer. Each ask contains information about the amount of resources that the user can contribute, the time window when it is available, and money remuneration. Upon receiving the asks, the service provider determines a completeness ratio, which is the ratio of the total resource from users and its resource demand. If the ratio is less than one, the service provider uses the storage and bandwidth from servers in data centers. The resource pooling cost is thus the sum of the payments to the users plus the marginal resource cost from the servers. The optimization problem for the service provider determines the allocation rule and payment to minimize the resource pooling cost, given the constraints ensuring the individual rationality of users and the truthfulness of the mechanism. To achieve the goals, the allocation rule and payment are designed according to a marginal pricing function, which is a non-increasing function of the completeness ratio. It was then proved that the allocation rule is monotone, and the online auction scheme is truthful. Simulation results indicated that the social cost in the online auction is always less than that of the offline VCG auction when varying the resource pooling demand, the number of asks, and the ratio between the server cost and the average asking price.

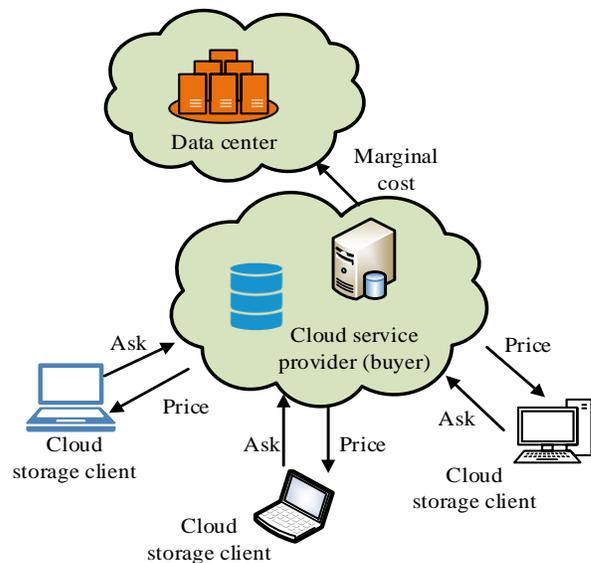


Fig. 18. Client-assisted cloud storage allocation using online reverse auction.

The client-assisted cloud model as in [226] still requires the communication among service providers and users. Therefore, when another user needs to use the resource, it makes a request

to the service provider which results in latency. Moreover, to guarantee the availability of resources, the interaction between demand users and supply users needs to be considered. Distributed cloud models such as self-organization cloud and social cloud can be adopted as discussed below.

2) *Self-organization cloud*: Self-organization cloud allows a number of host machines of users to be connected by a P2P overlay network on the Internet [228]. Since each user may act as a resource provider or a resource requester, resource exchange between them is typically modeled by the double auction. Such an approach was presented in [229] where the double auction is adopted for the task allocation among users. The model consists of request users, i.e., buyers, which require resources, i.e., computational resources and network bandwidth, for executing their tasks from provision users. Provision users act as sellers to contribute their resources for the task execution from the request. Before submitting bids, buyers have the rough estimation about the price of the required resources by using a price-setting mechanism, e.g., SpotCloud (<http://www.spotcloud.com>). The buyers then submit to a cloud planner, i.e., an auctioneer, their own bids including task descriptions, resource specifications, and the prices that they estimate. The cloud planner selects the buyer with the highest price as the winner. Then, the cloud planner sends the request of the winner to all sellers in the network. Interested sellers return the cloud planner by their asks. The seller with the lowest price is selected to provide resources to the buying winner. The lowest price is also the payment of the buying winner. However, how to discover users with available resource in the network was not given.

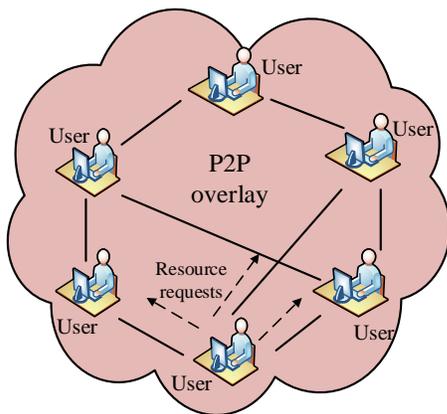


Fig. 19. A general system of self-organization cloud.

The authors in [230], [231] proposed to use the kademlia protocol [232] which allows a buyer to perform a *resource discovery* process to find candidate sellers with available resource in the network before the reverse auction is executed. Indeed, the candidate sellers can submit to the buyer their asks, each of which includes the information about available resources (i.e., CPU, memory, and bandwidth), QoS (i.e., a latency), a participation factor, and an incentive value. The participation factor is defined based on the historical resource contribution of a seller. The buyer calculates its utility value corresponding to each ask, taking into account weights assigned to each

component in the ask. The buyer selects a seller whose ask enables it to achieve the highest utility as the winner. The winner receives the incentive value representing the resources that it will receive in future. Such a non-money reward (along with coupons [233] or reputation score [234]) reduces the incentive costs for the buyer. This cost is significantly low as the number of sellers increases as shown in the simulation results. The resource allocation is thus fair and achieves the system stability. However, the reverse auction cannot be applied when the resource is insufficient.

3) *Social Cloud*: A social cloud is “a resource and service sharing framework utilizing relationships established between members of a social network” [235]. This model is thus similar to the self-organization cloud. However, if users in the self-organization cloud are anonymous and are not accountable for their actions, then accountability can be established through existing friend relationships in the social cloud.

The social marketplace is at the core of the social cloud which is similar to online procurement markets. Provision users in the social network arrive in a sequential manner to offer their services. Thus the posted-price mechanism is usually used. The authors in [5] adopted the posted-price model for storage service marketplaces in the social cloud as shown in Fig. 20. To enable the accountability, information of users such as user ID and credit balance, is managed by a bank. When each request user, i.e., a buyer, requests posted price offers for specific services, the cloud application checks the available balance of the buyer. The cloud application gives a list of provision users, i.e., sellers, their resource availability and the pricing information. These information is periodically updated and stored in a monitoring and discovery system. When the request user selects a service from a provision user, the cloud application creates SLA. If both parties accept the agreement, it is then passed to the bank for transferring credits between them. However, the provision users can lie about service costs to gain higher payoffs.

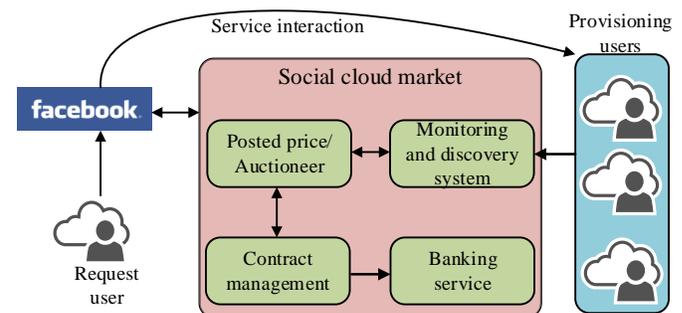


Fig. 20. Social cloud market infrastructure.

Therefore, the reverse Vickrey auction can be used as proposed in [235]. Upon receiving the storage service description of the request, an auction manager, i.e., an auctioneer, solicits asks from provision users within the social network. The asking prices are determined based on pre-defined metrics, and a seller with the lowest asking price is selected as the winner. Similar to [5], SLA between the buyer and the winner is created, and the credit transfer is done through a bank.

The credit is determined according to the payment policy of the reverse Vickrey auction which prevents misreports of provision users. The auction-based model is generally more complex than the posted-price model due to soliciting asks from sellers. To be appropriate in the social cloud environment with a large number of bidders, the auction manager needs to execute a number of concurrent auctions. As shown in the simulation results, a small scale auction manager may complete 65 concurrent auctions each with 50 bidders per minute.

To adapt to more requirements of users, the multi-attribute reverse Vickrey auction can be used. Multi-attribute auctions allow buyers and sellers to negotiate multiple attributes in addition to the price such as service quality and service deadline. The authors in [236] adopted such an auction for the task allocation among users in the social network. The request user initially sends its task description including a set of task attributes and the corresponding weights to a nearby auctioneer. The auctioneer broadcasts the request in the publishing area which may be described by the friends' community reachable through, e.g., a web-portal. Upon receiving the task request, potential provision users which meet the minimum attribute requirements submit their asks to the auctioneer. The auctioneer evaluates the utility score of each ask based on the available attributes of the ask and the request user's weights. The provision user with the highest utility score will be selected as the winner. The auctioneer then creates SLA involving the service price, the availability of resources, and the agreed attribute values between the winner and the buyer. The winner will get the price, which is increased by the amount, such that the utility of the transaction equals the utility score of the second-highest ask. However, how to define the request user's weights is not specified.

4) *Cloud-Centric Internet of Things*: This subsection considers an edge computing model in which the edge devices are sensing devices, e.g., smartphones of users, as shown in Fig. 13. Since these devices are equipped with various types of sensors, they are capable of generating sensing data. A large number of smartphones allow to gather sensing data from interest areas, and such a paradigm is called a crowdsensing network [237]. To provide the data to interested individuals/organizations, the crowdsensing networks and the cloud platforms are integrated. Since the crowdsensing network is also a main component of IoT [238], the integration can be called Cloud-centric Internet of Things (Cloud-IoT) [239] or Cloud of Things [240]. The main challenge in such models is to provide incentive mechanisms to users such that they perform sensing tasks and provide sensing data with the lowest cost. Typically, the reverse auction mechanisms are used.

The authors in [239] employed the reverse auction scheme to assign required tasks to phone users. The model has three entities: (i) cloud users which act as buyers, (ii) phone users, i.e., sellers, and (iii) a sensing server, i.e., an auctioneer. Once receiving a set of sensing tasks from the cloud users, the sensing server broadcasts the task requests to all phone users in a specific area. Interested phone users reply with asks, and the server sorts these users based on their marginal value contributions. The marginal value contribution of a user

is defined as an additional value introduced to the set by including the sensing task of the user [241]. Users with the highest marginal value contributions are selected for the tasks. The payments for the selected users are not less than their asking prices and determined based on their marginal value contributions to the set of sensing tasks.

In practice, mobility of phone users can degrade the utility of the platform since the users can move out a service region. The lightweight triangulation method [242] can be adopted for the server to keep track of the mobility pattern of each phone user. The utility of the platform also decreases due to the malicious phone users aiming at sending altered sensing data to the cloud users. Reputation of users is introduced to overcome this issue as proposed in [243] and [244]. The model and the reverse auction mechanism used in these approaches are similar to those in [239]. However, upon receiving sensing data from each phone user, the server runs an outlier detection (anomaly detection) algorithm [245] to classify the sensing data as an outlier or a non-outlier. Since the trustworthiness of a user can be time-varying, the user's instantaneous reputation is stored and updated in the database of the cloud platform. The current reputation of a user is determined as the function of the number of its non-outliers and outliers. When selecting winners, the server uses the ratio of its bidding price and current reputation to evaluate the bid. This selection process is to increase the selection probability of a phone user with a lower price and higher reputation, i.e., a smaller ratio. The payments for the winners are then applied similar to those in [239]. The simulation results showed that with the proposed approach, the payments of malicious phone users can be reduced around 55% compared with those in [239].

Summary: In this section, we have discussed four major models of the edge computing, and for each model we have reviewed the related economic and pricing approaches. We summarize the approaches along with references in Table IX and Table X. From the two tables, we observe that more client-assisted cloud systems have been recently studied. This is reasonable because the distributed cloud models reduce costs and service latency for users. In the next section, we review economic and pricing approaches for the resource management in cloud-based Video on Demand (VoD) systems. Cloud-based VoD systems are new video content delivery models in the development of cloud networking.

VII. APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD-BASED VOD SYSTEMS

This section describes and reviews the related work of cloud-based Video on Demand (cloud-based VoD), which is one important service supported by the cloud networking [246]. VoD, e.g., Internet Protocol TeleVision (IPTV) [247], is a system that enables users/clients to select and watch video contents whenever they want instead of watching at a specific broadcast time [248]. Traditional VoD services are based on the client-server or P2P architectures which have a major drawback of high costs and low bandwidth utilization, especially with the large and imbalanced demands of users.

TABLE IX
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN EDGE COMPUTING

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Cloudlet	[207]	Real-time group-buying auction	Cloudlet	Mobile users	Mobile videos	Seller forms a price curve of auction prices, and buyers with bidding prices higher than the auction price are the winners. The final deal price is the final auction price	Payment minimization, and expected profit maximization	Subgame perfect equilibrium
	[209]	Double auction	Cloudlets	Mobile users	Processing, storage, and networking	Based on sellers' asks, the auctioneer determines buyer and seller winning candidates. The buyer candidate with the highest price is the winner and is charged with a price of the second highest bid	Individual rationality, budget balance, and truthfulness	Market equilibrium
	[210]	Double auction	Cloudlets	Mobile users	Processing, storage, and networking	Winning buyer for each seller is determined based on sellers' asks. The clearing price charged to the winning buyer and the price paid to the winning seller are set at the same price	Individual rationality, budget balance, truthfulness, and system efficiency	Market equilibrium
	[72]	Non-cooperative game	Cloudlet	Brokers	Cloud resources	The Jacobi best-response algorithm is used to optimize the bidding prices of buyers	Cost minimization for mobile users	Nash equilibrium
	[213]	Supply and demand model	Mobile user	Mobile users	Cloudlet servers	Seller uses an aggregate excess demand function to define the total demand and total supply. The classical tatonnement process is used to adjust the price depending on the resource demand to clear the market resource	Payoff maximization, and resource efficiency	Market equilibrium
	[216]	Cost minimization	MTC brokerage	Mobile users	Cloud services	Given buyers' requests, the seller formulates the resource reservation as the total cost minimization. The linear programming and the min-cost greedy are used to solve the problem	Seller' profit maximization, and buyers' payment minimization	Optimal solution
Volunteer computing system	[219]	Reverse Vickrey auction	Volunteered computers	Task scheduler	CPU, memory, and network bandwidth	Buyer searches sellers with available resources via the distributed range query protocol. The buyer selects the seller with the lowest price as the winner, and the payment is based on the Vickrey auction policy	Payoff improvement, and truthfulness	Nash equilibrium
	[221]	Profit maximization	Volunteered computer owner	Users	CPU, memory, and network bandwidth	A broker determines the costs paid for the seller and the prices offered to the buyers to maximize the broker's profit. A sequential optimization is then introduced to solve this problem	Profit maximization	Optimal solution
	[222]	Demand-based pricing	Volunteered computer	Users	Task execution service	Seller assigns priority levels to tasks based on the k -means clustering algorithm and then executes tasks with the highest priority levels. The price is set based on the total resource demand of buyers	Throughput ration improvement, and payment reduction	Value optimization
	[224]	VCG auction	Bandwidth provider	Users	Bandwidth	Given buyers' requests, the seller finds an optimal schedule to maximize the sum of utilities of all buyers by using the greedy approach. The price is then set according to the VCG payment policy	Social welfare maximization	Nash equilibrium

Therefore, to reduce the costs and improve resource utilization, VoD providers, i.e., cloud tenants, can use cloud platforms from cloud providers to form the cloud-based VoD systems. The cloud-based VoD system can also provide flexibility to support users with various requirements [249]. However, the bandwidth cost is still significant in cloud-based VoD systems since video contents typically consume a large amount of bandwidth. The ultimate goal of VoD providers is to minimize the bandwidth cost and maximize their profit while satisfying users' demand, and thus pricing strategies are applied. In particular, the economic and pricing models have been used to address the following issues.

- *Bandwidth allocation*: In cloud-based VoD systems, VoD providers store their video contents in the cloud servers belonging to cloud providers. Therefore, the VoD providers need to reserve bandwidth which allows them to upload the video contents to the cloud servers as well as to guarantee the access for their users. Economic and pricing models have been used as solutions in which the bandwidth reservation cost is minimized while still satisfying users' demand.
- *P2P caching*: To reduce the reservation cost from the cloud servers, the VoD providers may use local resources,

e.g., the storage and upload bandwidth, of peers or users to cache video data. Since the users or peers are naturally selfish, pricing strategies have been adopted to incentivize the users to contribute their resources while minimizing the cost.

A. Cloud-Based VoD Models

This section presents economic and pricing approaches for bandwidth allocation in Cloud-based VoD systems. A major issue in such models is to allocate the bandwidth owned by cloud providers to VoD providers for delivering video contents. Depending on the specific scenario, a pricing model is applied to address the issue. In particular, if the model consists of a VoD provider and multiple cloud providers or multiple VoD providers and a cloud provider, competition-based pricing schemes such as auctions and non-cooperative game will be used. On the contrary, if the objective of the resource allocation is to maximize social welfare of all VoD providers, the network utility maximization can be adopted.

1) *Combinatorial auction*: The authors in [250] considered the cloud-based VoD system with a VoD provider and several cloud providers as shown in Fig. 21. The VoD provider delivers/allocates groups of videos from its local servers to

TABLE X
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN EDGE COMPUTING (CONT.)

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Client-assisted cloud systems	[226]	Online reverse auction	Cloud storage users	Storage service provider	Storage service	Based on sellers' asks, buyer uses a marginal pricing function to determine the allocation rule and the payment to minimize the resource pooling cost	Social cost minimization, truthfulness, and individual rationality	Nash equilibrium
	[229]	Double auction	Provision users	Request users	Resources	Given buyers' bids, a planner selects a buyer with the highest price and then finds a seller with the lowest price to provide resources to the buyer	Resource efficiency	Market equilibrium
	[230] [231]	Reverse auction	Provision users	Request user	Resources	Based on sellers' asks, the buyer calculates the corresponding utility values. The seller whose ask maximizes the buyer' utility is selected as the winner. The winner then gets an incentive value	System stability, and low incentive cost	Nash equilibrium
	[5]	Posted-price	Provision users	Request user	Storage service	Based on a list of sellers along with their posted price offers, the buyer selects a seller and the cloud application creates an SLA. The buyer then pays the seller through the bank	Utility maximization for buyer	Nash equilibrium
	[235]	Reverse Vickrey auction	Provision users	Request user	Storage service	Given sellers' asks, the auction manager selects a seller with the lowest price as the winner. The payment follows the payment policy of the reverse Vickrey auction	Truthfulness, and payment minimization	Nash equilibrium
	[236]	Multi-attribute reverse Vickrey auction	Provision users	Request user	Task execution service	The auctioneer evaluates the utility scores corresponding to sellers' asks. The seller with the highest utility score is selected as the winner. The payment is determined based on the payment policy of the reverse Vickrey auction	Utility maximization of buyer, truthfulness	Nash equilibrium
	[239]	Reverse auction	Phone users	Cloud users	Sensing task	The server selects the sellers with the highest marginal contributions as the winners. Payments for the winners are determined based on their marginal contributions	Incentive cost minimization	Nash equilibrium
	[243] [244]	Reverse auction	Phone users	Cloud users	Sensing task	Same as [239], but for each seller, the server defines a ratio of its asking price and reputation. The reputation is determined using the outlier detection algorithm. Sellers with smaller ratios are selected as the winners	Utility maximization, and incentive cost reduction	Nash equilibrium

the cloud providers. Since there are several combinations of videos for trading, the combinatorial auctions are adopted. The VoD provider as a buyer classifies videos into groups, and each group consists of the same user demands. The VoD provider also evaluates price for each group, which is generally a decreasing function of user demands. Then, the VoD provider sends the requests to the cloud providers. Depending on available bandwidth and memory, each cloud provider, i.e., a seller, determines the number of groups and the number of videos in each group that it can serve. Cloud providers respond to the VoD provider with their asks including the information related to the number of groups and videos along with the corresponding prices. The VoD provider computes a *distance* value associated with each ask. The distance value is the price evaluated by the VoD provider minus the price offered by the cloud provider. The VoD provider selects a cloud provider with the largest distance value as the winner to allocate its video groups. To guarantee the truthfulness of the auction, the payment policy from the Vickrey auction is adopted. The Approximate Efficiency Maximization (AEM) algorithm [251] is used to avoid the collusion among cloud providers. The simulation results showed that the proposed approach saves up to 10% of the cost compared to the video migration strategy in [252]. However, multiple VoD providers need to be considered in the future work.

2) *Generic pricing mechanism*: The authors in [253] extended the market model in [250] involving multiple public cloud providers, i.e., sellers, VoD providers (cloud tenants), i.e., buyers, and a broker. In this setting, the broker reserves the actual bandwidth from the cloud providers and sells

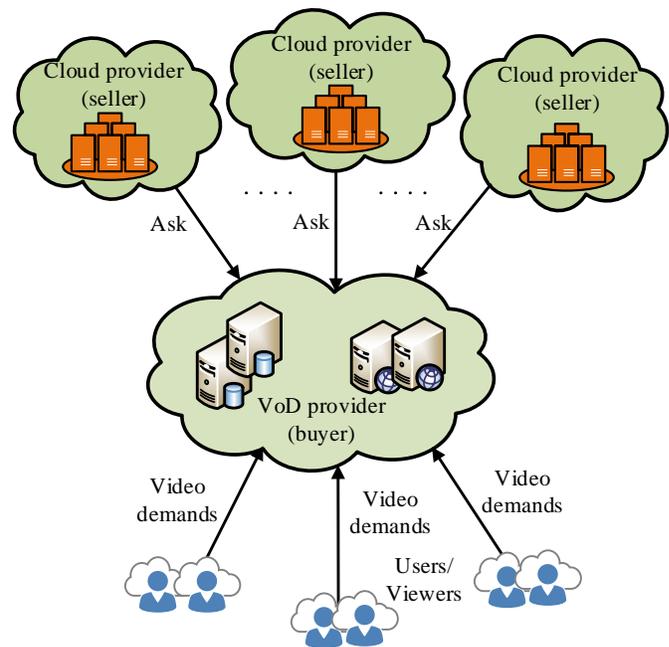


Fig. 21. Bandwidth allocation in cloud-VoD system based on combinatorial auction.

probabilistic bandwidth guarantee services to tenants. The broker determines the lower and upper bounds of the price to maximize its profit and minimize the bandwidth reservation cost. The broker then determines a *load direction matrix* which directs tenants' bandwidth demands to the cloud providers. To

TABLE XI
A SUMMARY OF ADVANTAGES AND DISADVANTAGES OF MAJOR APPROACHES FOR THE RESOURCE MANAGEMENT IN EDGE COMPUTING.

Major approaches	Advantages	Disadvantages
[207]	<ul style="list-style-type: none"> • Achieve good economic properties 	<ul style="list-style-type: none"> • Have slow convergence • Have unstable equilibrium
[210]	<ul style="list-style-type: none"> • Reduce communication latency • Improve the system efficiency 	<ul style="list-style-type: none"> • Have slow convergence • Have unstable equilibrium
[72]	<ul style="list-style-type: none"> • Support both long-term reservation and on-demand request 	<ul style="list-style-type: none"> • Require buyers to submit their requests simultaneously • Have unstable equilibrium
[221]	<ul style="list-style-type: none"> • Guarantee reliability to buyers 	<ul style="list-style-type: none"> • Be challenging to learn probabilities of price acceptance for buyers • Have high computational complexity
[226]	<ul style="list-style-type: none"> • Support asynchronous arrivals of both buyers and sellers 	<ul style="list-style-type: none"> • Require high communication
[5]	<ul style="list-style-type: none"> • Support anonymous buyers • Achieve high reliability and resource availability 	<ul style="list-style-type: none"> • Require frequently monitoring and discovering the resource availability and prices of sellers
[239]	<ul style="list-style-type: none"> • Support the mobility of sellers 	<ul style="list-style-type: none"> • Require frequently calculating and updating instantaneous reputation of sellers

find the lower bound of the price, the broker defines its profit which is the sum of prices offered to tenants minus the total cost for reserving bandwidth from the cloud providers. The price offered to the tenant is a concave function of its demand, and the demand is assumed to follow the Gaussian distribution [254]. The lower bound of the price is determined by using the gradient ascent algorithm for the profit maximization. For the upper bound of the price, the broker needs to set the price lower than that offered by the cloud providers to attract more tenants. Therefore, the upper bound of the price is actually the cloud providers' pricing scheme [255]. The simulation results showed that the broker can save the bandwidth reservation cost by more than 30% on average compared with the case that each tenant reserves bandwidth individually.

3) *Non-cooperative game*: For the market model in [253], the pricing scheme of the broker is affected by the cloud providers, and such a market is considered to be a *controlled market*. The authors in [256] considered a *free market* which only consists of tenants, i.e., VoD providers and a broker. The tenants competes with each other for the cloud bandwidth by submitting their pricing strategy to the broker. The interactions among selfish tenants are modeled as a non-cooperative game in which the strategy of each tenant is to set the price so as to maximize its own utility. The utility is inversely proportional to the price that the tenant pays. Using the Cauchy-Schwarz inequality and the proof by contradiction, it was shown analytically that if the broker decides the *load direction matrix* to maximize its profit as mentioned in [253], the tenants' prices will converge to a unique Nash equilibrium. This equilibrium still holds even if multiple brokers exist in the market since the game is played by the tenants. However, the competition among brokers may lead to a zero profit of brokers.

4) *Utility maximization*: Unlike the above market models, the authors in [257] and [258] considered multiple VoD providers (cloud tenants), which reserve the egress network bandwidth, i.e., the upload speed, from a data center of a cloud

provider to guarantee delivering smoothly videos to their users. The objective is to find the resource allocation that maximizes the tenants' social welfare. The social welfare is the sum of tenants' utilities minus the total cost at the cloud provider, and thus the utility maximization problem as presented in Section III-C can be applied. In particular, the tenant utility is a strictly concave and monotonically increasing function of the allocated resource while the total cost is a strictly convex function that is also monotonically increasing with the allocated resource. Such concavity of the utility function is to guarantee that the optimization problem has a unique optimal solution. Since the utility function may not be known by the cloud provider, and the total cost function may not be known to the tenants, the centralized algorithms, e.g., the Newton's method [120], cannot be applied. The authors in [257] and [258] adopted two distributed iterative algorithms based on price updates.

In [257], at each iteration, the cloud provider updates the price charged to each tenant in the next iteration using the first-order partial derivative of the total cost with respect to the tenant's resource allocation and the price at the current iteration. Given the price at the current iteration, each tenant determines its resource allocation so as to maximize its utility minus the cost of using resources. Unlike [257], the price in [258] is updated by the tenants while the resource allocation is updated by the cloud provider. At each iteration, each tenant sets the price in the next iteration based on the first-order derivative of its utility function and the price at the current iteration. In both [257] and [258], it was proved that if the price and the resource allocation reach fixed points, then the resource allocation will be the optimal solution of the optimization problem.

In general, compared with the algorithms such as the Newton's method and the Alternating Direction Method of Multiplier (ADMM) [259], the proposed approaches minimize the message passing overhead. This is because the approaches use only their local information, i.e., the cost and utility functions,

and feedback update variables. Moreover, the approaches remove the assumption on the cooperativeness of entities in the market. For example, in [257] once the cloud provider sets a price vector, the tenants will find their own resource allocation to maximize their utilities, which are natural strategies of selfish tenants. This property is in contrast to the ADMM [259] which requires more complicated variable updates. The simulation results showed that the average number of iterations needed to converge of the proposed approach in [257] is 10 while those of the primal gradient descent algorithm and the gradient descent-based consistency pricing method [260] are 100 and 50, respectively. However, a more general market with multiple cloud providers needs to be investigated in the future work.

B. P2P-Assisted Cloud-Based VoD Models

As stated earlier, an important challenge in cloud-based VoD systems is the bandwidth cost on the servers in the cloud. For example, YouTube is estimated to spend \$470 million a year while Facebook spends \$500,000 a month on bandwidth (<http://www.slate.com/articles/technology/technology/2009/>). To improve scalability and save bandwidth costs at the cloud, the cloud-based VoD systems can be combined with P2P networks as shown in Fig. 22. Peers, i.e., users, can download video contents from both the cloud and other peers in the P2P network. However, this requires the VoD provider to provide the peers with an incentive for (i) downloading videos from other peers rather than the cloud and (ii) caching the video contents by using their resources, e.g., the memory and upload bandwidth. Pricing models were introduced to achieve these goals as discussed in what follows.

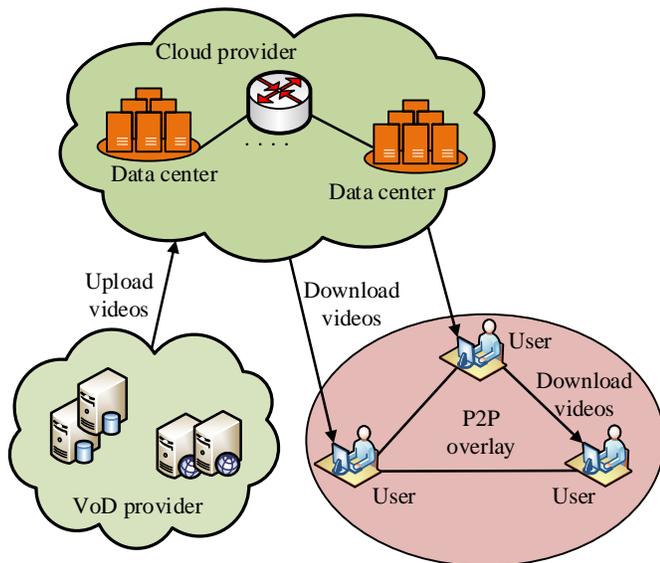


Fig. 22. Overview of a P2P-assisted cloud-based VoD model.

1) *Stackelberg game*: Typically, to reduce the access of users to servers in the cloud, the VoD provider can apply smart pricing strategies, e.g., the usage-based pricing, to the bandwidth consumption. To optimize both utilities of the VoD provider and users, the Stackelberg game can be adopted as

proposed in [78]. First, the VoD provider which acts as the leader estimates the cloud bandwidth usage of users by using the exponentially weighted moving average method [261]. Then, it sets the VoD service price proportionally to the estimated cloud bandwidth usage. Given the price, users, i.e., followers, select bit rates to maximize their utility. The utility is the difference between the satisfaction degree and the price that the user pays to the VoD provider. The satisfaction degree is a concave function of the bit rate. Upon receiving users' responses, the VoD provider sets the price so as to maximize its revenue. Based on the new price, each user recalculates the optimal bit rate that maximizes its utility. The simulation results showed that the cloud bandwidth consumption of the proposed approach is significantly lower than that of the server-side bit rate adaptation approach in [262]. The reason is that the proposed approach encourages users to download video chunks from peers through the usage-based pricing. However, how the users determine their optimal bit rates was not specified.

2) *Double auction*: Another solution which reduces cloud bandwidth consumption is to share the surplus upload bandwidth among peers, called the prefetching strategy [263]. As mentioned in Section VI-C2, the resource sharing among peers is typically implemented in a decentralized manner since any peer may be a buyer or a seller. Thus, the resource sharing can be modeled as the double auction-based market as proposed in [264]. The market is divided into several sub-markets, each of which only trades one video segment. When a peer acts as a buyer, it broadcasts a bid to its neighbors. A seller compares its ask with the bid. If the ask is less than the bid, the seller will propose the buyer to trade the segment at a transaction price, which is the average of the ask and the bid. Since the transaction price is less than the bid and larger than the ask, the proposed approach is the ex-post individual rationality, i.e., the expected utility of participants is non-negative. This incentivizes peers to exchange video segments with each other rather than to download from the server. The simulation results showed that the proposed approach outperforms the near-sequential prefetching strategy [265] and the popularity-based prefetching strategy [266] in terms of server bandwidth costs. However, since the buyer only receives a video segment in a sub-market, it may need to participate in other sub-markets to satisfy its demand. This increases the latency of delivering the video.

3) *Generic pricing mechanisms for the P2P caching*: In practice, peers may not be satisfied due to the lack of some video contents in the market. One possible reason is that the other peers may clear their local storage after finishing watching the videos. The authors in [267] designed a reward price-based incentive mechanism for peers to cache the video replicas to satisfy all peers' demand. Videos in the system are categorized into different classes, each consisting of videos of similar popularity associated with a price. The VoD provider determines these prices such that the number of supplied video replicas equals the demand. First, it defines the *refreshing probability* of each peer, i.e., the probability that the peer clears its local storage. Second, it defines the *storage probability* which can be considered to be the *popularity* of a video. Since

each peer decides whether to cache a video based on the price of the video, the price of each video is proportional to the refreshing probability, the desired number of replicas of the video in the system, and inversely proportional to the storage probability. For example, videos which are more popular have higher prices to incentivize more peers to cache them so as to meet the greedy cache requirement [268].

In reality, the VoD provider always considers its profit when setting prices for videos. The authors in [269] addressed this problem by investigating *strategic pricing* to minimize the VoD provider's operation cost. The VoD provider calculates the operation cost involving the upload cost to the cloud and the reward price paid to peers. The upload cost is proportional to the difference amount between the desired videos and the supplied videos. The reward price is the total cost paid to all peers due to their video contributions. The operation cost is a continuous and convex function of the reward cost. Thus, given a budget for reward prices, it was proved that there exists a unique solution of the reward price to the cost minimization problem. The simulation results revealed that compared with the approaches without using any incentive scheme, the proposed approach reduces significantly the operation cost, especially when the upload cost of servers increases. In practice, there is a fraction of peers which may not be sensitive to the reward prices. Thus, the cost may be reduced further if the VoD provider can learn the real sensitivity of the peers.

4) *Pricing models for resource allocation in cloud-assisted P2P streaming systems*: This section reviews a few pricing approaches to provide incentive to peers in cloud-assisted P2P streaming systems. Cloud-assisted P2P streaming systems are similar to the P2P-assisted cloud-based VoD models [270]. However, in the cloud-assisted P2P streaming systems, there is no specific VoD provider, and thus stakeholders involve only peers and streaming servers in the cloud.

As presented in [71], the cloud as a buyer rents resources including network bandwidth, storage space, and CPU, from peers as sellers, namely *helpers*, to contribute video contents to its customers. The cloud offers the service price, and then the peers decide their resource contributions. Thus, the Stackelberg game can be used. Each peer, i.e., a follower, computes its payoff which is the difference between the price offered by the cloud, i.e., the leader, and the cost incurred to offer the resource. The strategy of the peers is to find the amount of resources to maximize their own payoffs. The optimal amount can be determined by using the first-order derivative. Given the total amount of the resources from peers, the cloud determines the offered price to maximize its revenue.

The approach in [71] did not consider the budget of the cloud. The authors in [271] adopted the Stackelberg game to analyze the upload bandwidth sharing between helpers and a streaming server resided in the cloud, taking into account the budget of the server. In the first stage, the server, i.e., the leader, announces its budget. The second stage can be considered to be a non-cooperative game among selfish helpers, i.e., followers, which decide the number of bandwidth units to maximize their own utilities. The utility is the difference between the reward that the helper receives and its video sharing

cost. Generally, the utility is a strictly concave function of the number of bandwidth units. By using the second derivative, it was proved that there exists a unique Nash equilibrium involving optimal strategies of helpers. Given the helpers' strategies, the server determines the budget to maximize its utility, which is the gain from the total transmitted video minus the reward paid to the helpers. In particular, the gain is characterized by the two-parameter rate-distortion model which expresses the trade-off between the allocated bandwidth and the video content distortion. The utility is a strictly concave function of the budget. It was also proved that there is a unique budget value which maximizes the utility.

The server can determine an optimal value of budget only if it has full information about the helpers' utility functions. In real scenarios, this assumption may not be valid since the helpers autonomously decide their upload bandwidth contributions. Thus, the reverse auction can be used. The server, i.e., the buyer, broadcasts a vector of bandwidth demands to helpers, i.e., sellers. Interested helpers submit to the server their asks including the number of bandwidth units and the prices that they are willing to pay. Given the asks, the server computes its utility. The utility is the difference between the value that the server would pay if it did not receive helps from helpers and the sum of asking prices. The server needs to select a set of winners to maximize the utility within its budget. The optimization problem is NP-hard. However, the server's utility was proved to have the submodularity property. Thus, the winner determination and the payment rules can be implemented by using the results on the budget feasible mechanism design in [272] for submodular functions. The experiment results showed that the reverse auction-based approach outperforms the greedy algorithm in terms of server utility and truthfulness. However, the number of shared bandwidth units achieved in the reverse auction-based is less than that of the Stackelberg game-based approach, which can access to full information about the helper utility function.

C. Cloud-based Wireless Multimedia Social Networks (CWMSN)

Cloud-based Wireless Multimedia Social Networks (CWMSN) are proposed to support heterogeneous services to users. The CWMSN model is shown in Fig. 23 which is essentially a combination of the multimedia cloud and different subnetworks. These subnetworks are based on the various social contexts, e.g., family, interest, and hobby. The system includes content providers, desktop users, and mobile users. The content providers deliver their live programs such as live-streaming and VoD, video conferences, VoIP, photo sharing and editing to the desktop users through distributed servers and gateways. Desktop users have higher computational capability and more resources than those of the mobile users. Therefore, desktop users can share their resources, i.e., the bandwidth, with mobile users which want to obtain live programs. Since the mobile users are selfish for resource competition, game theory-based pricing schemes are appropriate solutions to analyze the bandwidth sharing.

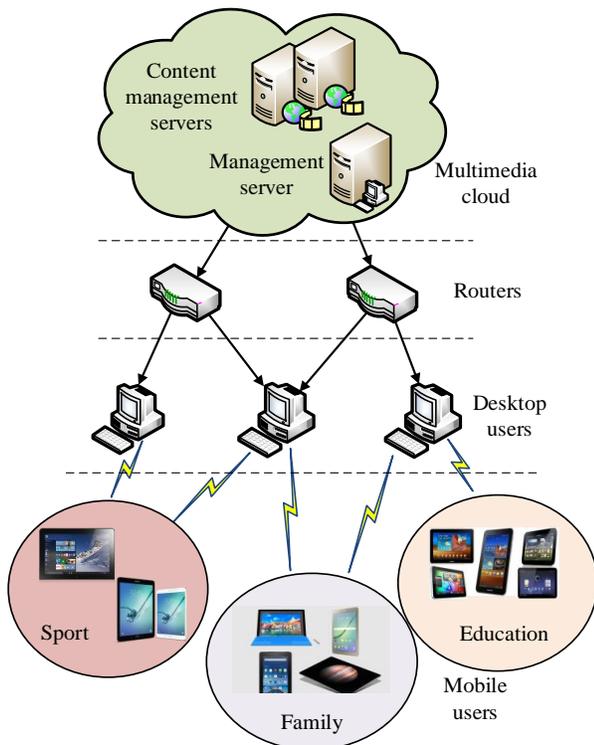


Fig. 23. Cloud-based multimedia social network architecture.

1) *Stackelberg game*: In the context of CWMSN, the authors in [273] addressed the bandwidth sharing issue between the desktop users and the mobile users. The desktop users decide the amount of bandwidth and the corresponding price, and then each mobile user selects the specific desktop user to connect to. The interactions between them can be formulated as the Stackelberg game in which the desktop users are leaders and the mobile users are followers. Mobile users are not fully rational when making their decisions, and their behaviors are thus modeled by the evolutionary game. Initially, each mobile user in a group randomly connects to a desktop user and calculates its utility based on the allocated bandwidth and the price offered by the desktop user. The mobile user also computes the average utility of the group. If the average utility is greater than its own utility, the mobile user changes its connection to another desktop user to possibly receive higher utility. Otherwise, it keeps the current connection. This process is repeated until all mobile users in the same group achieve equal utility at the equilibrium. Given the result of the mobile users' evolution, desktop users compete with each other in a non-cooperative game by deciding the amounts of bandwidth and corresponding prices to maximize its utility. The solution of the non-cooperative game is the Nash equilibrium which is computed by finding a fixed point of the best response functions of all desktop users [274]. However, the proof for the existence and uniqueness of the Nash equilibrium was not given.

The bandwidth allocation in [273] is the proportional allocation mechanism, meaning that different mobile users receive the same amount of bandwidth if the ratio of their bids to

their minimum bandwidth requirements is a constant. This is unfair to the desktop users since the mobile users may pay less while they still receive the requested amount of bandwidth, as long as the ratio is constant. The authors in [275] introduced a punishment coefficient to evaluate the bandwidth of the mobile user obtained from the desktop user. The punishment coefficient is the ratio of a mobile user's bid and its minimum bandwidth requirement. The mobile user is considered to have a cheating behavior if its punishment coefficient is less than one since its bid cannot be lower than its minimum bandwidth requirement. Then, the bandwidth obtained by the mobile user will be reduced by the corresponding punishment. The simulation results highlighted that with the proposed solution, i.e., the cheat-proof strategy, each mobile user must honestly bid the amount of bandwidth which is equal to its minimum bandwidth requirement to obtain its desired amount. By contrast, with the non-cheat-proof method, the mobile users can obtain the same amount of bandwidth, but they only pay less.

In practice, the Nash equilibrium as mentioned in [273] and [275] cannot be obtained if the desktop users do not have the information about each other's strategies, i.e., the amount of shared bandwidth and the price. Learning-based algorithms that adjust strategies toward increasing the utility of the desktop users can be used to update this information.

2) *Generic pricing mechanism*: The approaches in [273] and [275] did not consider the characteristic of mobile users. Using the same model, on the contrary, the authors in [276] divided the mobile users into two categories which are price-sensitive users and QoS-sensitive users. Each desktop user determines its two pseudo-demand functions corresponding to these two types of mobile users. In general, these functions depend on the possible maximum number of mobile users and the bandwidth price. The desktop user formulates its benefit maximization problem based on the demand functions. The problem is then solved by using the first-order derivative and the Girolamo Cardano algorithm [277] which provides the optimal price to each desktop user. Since the number of mobile users connecting to a specific desktop user may vary in each time period due to price changes by other desktop users, the desktop user needs to continuously adjust the bandwidth price to maximize its utility. The desktop users may change their bandwidth prices until their total non-decreasing utility is unchanged. However, the price adjustments did not consider the constraints on the available bandwidth of the desktop users.

Summary: In this section, we have discussed two major issues in cloud-based VoD systems. We have reviewed economic and pricing approaches for these issues. A summary of these approaches is given in Table XII. As shown in the table, most approaches address the bandwidth allocation issue in the cloud-based VoD systems since the bandwidth consumption for the video distribution in these systems is crucial. Moreover, the Stackelberg game is considered to be an efficient pricing model in guaranteeing utility maximization for both sellers and buyers. In the next section, we review economic and pricing approaches for the resource management in cloud-based Software Defined Wireless Network (SDWN) model. SDWN is a combination of the cloud (i.e., the cloud data

TABLE XII
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD-BASED VOD SYSTEMS.

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Cloud-based VoD system	[250]	Combinatorial auction	Cloud providers	VoD provider	Video instances	Given sellers' asks, the buyer selects the winner based on the difference between the price evaluated by the buyer minus sellers' asking prices. The payment is based on the Vickrey auction	Cost minimization, and truthfulness	Optimal solution
	[253]	Generic pricing	Public cloud providers	VoD providers	Probabilistic bandwidth guarantees	A broker determines the bounds of price offered to each buyer based on the Gaussian distribution of the buyer's demand and the pricing policy of the sellers	Bandwidth reservation cost minimization, and cloud resource efficiency optimization	Nash equilibrium
	[256]	Non-cooperative game	Broker	VoD providers	Cloud bandwidth	Buyers submit their pricing strategies, and optimal prices are achieved using the Cauchy-Schwarz inequality and the proof by contradiction	Utility maximization for buyers	Nash equilibrium
	[257]	Utility maximization	Cloud provider	VoD providers	Egress network bandwidth	Seller updates the price using the first-order partial derivative of its total cost. Then, the buyer determines its optimal resource allocation based on the updated price	Social welfare maximization, and profit maximization	Optimal solution
	[258]	Utility maximization	Cloud provider	VoD providers	Egress network bandwidth	Buyers update the price using the first-order derivative of their utility functions. The seller determines the optimal resource allocation for buyers	Social welfare maximization, and profit maximization	Optimal solution
P2P-assisted cloud-based VoD system	[78]	Stackelberg game	Cloud provider	Users	Bandwidth	The VoD provider sets the price using the usage-based pricing. Buyers select bit rates to maximize their utilities, and then the VoD provider sets the price so as to maximize its revenue. Each buyer recalculates the optimal bit rate	Optimal bit rate for buyers, and revenue maximization for the seller	Stackelberg equilibrium
	[264]	Double auction	Peers	Peers	Video segment	If an ask of a seller is less than a bid of a buyer, there is a transaction at which the price is the average of the ask and the bid	Ex-post individual rationality, budget balance, and cloud bandwidth reduction	Market equilibrium
	[267]	Generic pricing	Peers	VoD provider	Video caching service	Buyer sets the price for caching video depending on the refreshing probability, the desired number of replicas of the video in the system, and the storage probability of the video	Ex-post individual rationality, budget balance, and cloud bandwidth reduction	Market equilibrium
	[269]	Cost minimization	Peers	VoD provider	Video caching service	Buyer calculates its operation cost involving its upload cost to the cloud and the reward price paid to peers. The reward price is the solution of the operation cost minimization problem	Operation cost minimization	Optimal solution
	[71]	Stackelberg game	Peers	Cloud	Network bandwidth, storage space, and CPU power	Buyer offers the price for contributing its video contents, and then sellers decide their optimal resource contributions. Given the total resource contribution, the buyer determines the offered price by using the first derivative	Payoff maximization for sellers, and utility maximization for buyer	Stackelberg equilibrium
	[271]	Stackelberg game	Peers	Streaming server	Upload bandwidth	Sellers decide the number of bandwidth units based on the buyer's budget. Then, the buyer determines the optimal budget using the first-order derivative	Utility maximization for sellers and buyer	Stackelberg equilibrium
	[271]	Reverse auction	Peers	Streaming server	Upload bandwidth	The buyer computes its utility based on sellers' asks. The winner determination and the payment rules are implemented based on the budget feasible mechanism design for submodular functions	Truthfulness, and utility maximization for buyer	Nash equilibrium
CWMSN	[273]	Stackelberg game	Desktop users	Mobile users	Bandwidth	Each buyer in a group will change the current connection to the seller until all buyers in the group achieve the equal utility. Then, sellers compete with each other by deciding the amounts of bandwidth and the corresponding prices	Equal utilities for buyers, and payoff maximization for sellers	Stackelberg equilibrium
	[275]	Stackelberg game	Desktop users	Mobile users	Bandwidth	Same as [273], but a punishment coefficient is introduced to avoid the cheating behavior of buyers. The coefficient is the ratio of a buyer's bid and its minimum bandwidth requirement	Equal utilities for buyers, payoff maximization and fairness for sellers	Stackelberg equilibrium
	[276]	Generic pricing	Desktop users	Mobile users	Bandwidth	Each seller determines the pseudo-demand function of the bandwidth demand of buyers and then formulates its benefit maximization problem	Utility maximization for sellers, and efficient allocation	Utility equilibrium

center networking), Software-Defined Networking (SDN), and wireless networks.

VIII. APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD-BASED SDWN

Distributed data centers in cloud data center networking as discussed in Section IV can reduce data transfer cost and delay for users. However, the geo-distributed networks increase the difficulty of global resource management unless there is a centralized control. SDN (see Section II-B4) provides a real-time centralized control based on both instantaneous network status and user defined policies. In practice, SDN has been adopted in wireless networks to form the Software

Defined Wireless Network (SDWN) [278], [279], [280] for the centralized control and global optimization [281]. Thus, data center networking can be combined with SDWN, namely cloud-based SDWN, for complex network management as illustrated in Fig. 24. In the network model, the SDN controller acts as a "brain" of the network. It monitors and allocates resources from data centers to users via wireless networks, i.e., cellular networks and WiFi hotspots. Such centralized resource management usually aims at optimizing the total benefit of all stakeholders. Thus, economic and pricing models such as bargaining game, network utility maximization, and stackelberg game, are appropriate solutions since their outcomes can guarantee maximizing the overall utilities of all stakeholders.

TABLE XIII

A SUMMARY OF ADVANTAGES AND DISADVANTAGES OF MAJOR APPROACHES FOR THE RESOURCE MANAGEMENT IN CLOUD-BASED VoD SYSTEMS.

Major approaches	Advantages	Disadvantages
[250]	<ul style="list-style-type: none"> • Achieve economic efficiency • Avoid collusion 	<ul style="list-style-type: none"> • Have high computational complexity • Support only one VoD provider
[258]	<ul style="list-style-type: none"> • Support multiple VoD providers • Minimize communication overhead 	<ul style="list-style-type: none"> • Have slow convergence • Support only one cloud provider
[78]	<ul style="list-style-type: none"> • Have stable equilibrium • Achieve win-win solution 	<ul style="list-style-type: none"> • Have slow convergence
[269]	<ul style="list-style-type: none"> • Adaptive to nonasymptotic system and highly dynamic video popularity 	<ul style="list-style-type: none"> • Be challenging to learn the feature of system dynamics and adjust the pricing scheme
[271]	<ul style="list-style-type: none"> • Achieve computational efficiency 	<ul style="list-style-type: none"> • Do not support online situations, i.e., the available bandwidth changes dynamically
[275]	<ul style="list-style-type: none"> • Prevent cheating behaviors 	<ul style="list-style-type: none"> • Have unstable equilibrium • Require having information about each other's strategies

In particular, the economic and pricing models have been used to address the following issues.

- *Bandwidth allocation*: Bandwidth allocation in the cloud-based SDWN is to allocate bandwidth from data centers owned by cloud providers to Service Providers (SPs) and mobile users in a centralized manner at the SDN controller. Economic and pricing models have been used to maximize the payoffs for cloud providers, service providers, and mobile users simultaneously.
- *Mobile data offloading*: Mobile data offloading, also known as WiFi offloading, is the use of complementary network technologies to reduce the amount of data being transferred through cellular networks. In the cloud-based SDWN, the mobile data offloading is enabled by SDN at an edge of the network to dynamically position or reposition the traffic. However, since the complementary networks and cellular networks may belong to different parties, the traffic offloading is implementable only if the benefits of the parties are satisfied. Economic and pricing models such as the contract theory have been adopted to guarantee the condition.

Before discussing these approaches, the cost analysis is introduced to evaluate the cost of wireless networks when SDN is enabled.

A. Cost analysis

The authors in [55] analyzed the costs of the LTE network owned by a Mobile Network Operator (MNO) with and without SDN, denoted as SDN-LTE and non-SDN-LTE, respectively. The Finnish mobile network topology was adopted as a reference model. The costs generally are divided into the CAPital EXpenditure (CAPEX) and OPerational EXpenditure (OPEX). CAPEX involves costs of network equipments and their deployment cost. The SDN-LTE configuration is implemented in a more centralized manner, and thus the deployment cost is lower than that of non-SDN-LTE. OPEX includes energy consumption, site visits, and network management expenses. Since SDN increases the automation of fault detection,

the network management expenses in SDN-LTE are expected to decrease compared with those in non-SDN-LTE. Moreover, the site visit costs in the non-SDN-LTE model arise since base stations, i.e., eNBs, and switches are distributed across the country which may incur more travel expenses and time on site. The quantitative results showed that SDN-LTE can reduce the annual CAPEX by around 7.72% and the annual OPEX by 0.31% compared with non-SDN LTE. However, these savings are relatively small compared with the total annual cost of MNO.

Similarly, the authors in [56] evaluated the total cost of the LTE network through CAPEX and OPEX when the Network Function Virtualization (NFV)/SDN approaches are implemented. Three possible scenarios were considered. The first scenario is that MNO owns the Virtualized Network Functions (VNFs)/SDN and hardware. CAPEX involves initial investments of the VNF/SDN software licences and the hardware. OPEX includes the maintenance cost, variable costs such as energy consumption and cooling cost, software update, certificate update and bug fixing costs. In the second scenario, MNO owns VNF/SDN and rents hardware. CAPEX is still the VNF/SDN software licenses, but OPEX is the annual rental fee for the required hardware. In the third scenario, MNO rents both VNF/SDN and hardware. In this case, there is no CAPEX while OPEX involves the cost incurred for VNF/SDN, annual service charge, e.g., the hardware and software maintenance, and energy consumption. The simulation showed that the total cost when VNF/SDN are used is always lower than that without VNF/SDN. Moreover, the total cost in the second scenario is the lowest over the period from 2014 to 2019.

The evaluations using the above cost model show the benefit in terms of cost minimization when SDN is enabled. In the following, we provide some other economic and pricing models for the SDN-based resource management in the cloud-based SDWN. Note that although most approaches in this section addressed the issues in the cloud-based SDWN, a few wireline-enabled SDN models, e.g., SDN-enabled home networks, will be also discussed as an extension.

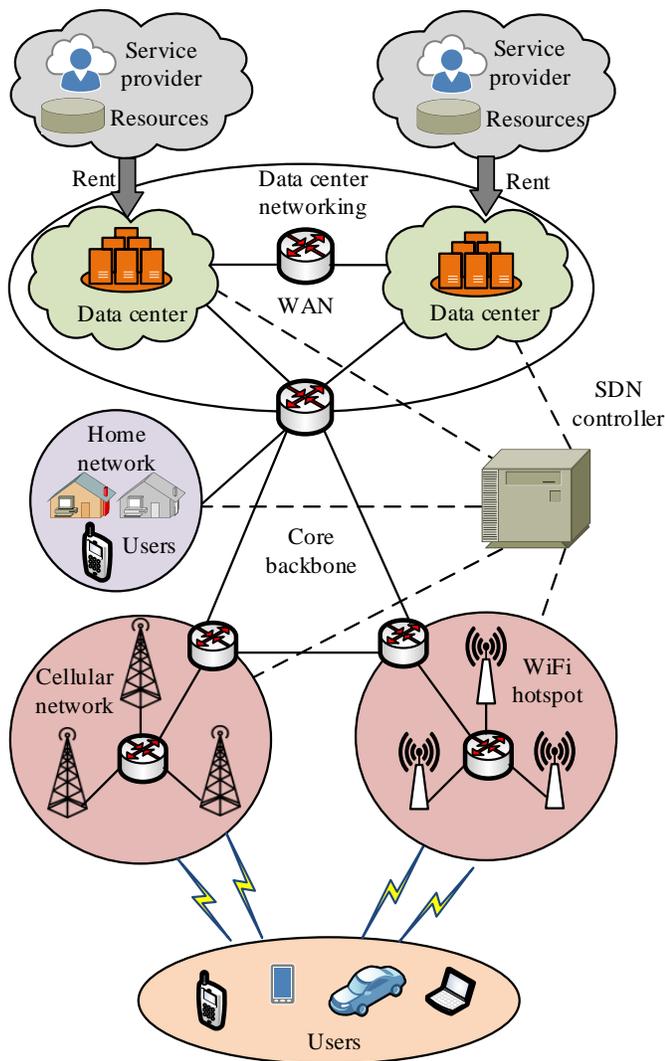


Fig. 24. An illustration of cloud-based software-defined wireless network.

B. Bargaining Game for Bandwidth Allocation

Considering the cloud-based SDWN, the authors in [86] investigated the issue of sharing resources, i.e., CPU, memory and bandwidth, among service providers to extend their capabilities in serving mobile users. The service providers can be divided into sellers and buyers. The main objective is to guarantee QoS for users while maximizing the total increased utility of all sellers and buyers. Thus, the Nash bargaining game with cooperation policy was adopted. The sellers cooperate with each other to form a resource pool, and the buyers bid for the resources in the resource pool with a unit price. The increased utility of the seller is obtained from leasing resources and the price while the increased utility of the buyer is defined by the difference of its utility after and before renting resources from the sellers. The objective function is the total increased utility of all buyers and sellers. The deal price is then determined based on the Cardano's formula [282], and the basic optimization is used to obtain the total allocated resource. The buyers share the total resource corresponding to their demands. The simulation results

showed that the proposed approach increases both allocated resources for buyers and revenue for sellers compared with the bargaining game in which buyers compete with each other. However, the bargaining game with the competition policy is more appropriate in the realistic scenarios due to the resource scarcity.

C. Pricing Models for Mobile Data Offloading

As shown in Fig. 24, Access Points (APs) can be integrated in the cloud-based SDWN to minimize the cloud bandwidth consumption for the base stations [283], and thus reducing the service cost and latency for mobile users. In this context, two economic and pricing models were adopted corresponding to two specific scenarios. In the first scenario, the access points and base stations are owned by one service provider, and the aim of the mobile data offloading is to maximize utility functions of all mobile users. Thus, the NUM framework was used. In the second scenario, the access points and base stations are owned by different providers, and the access points will only admit the traffic from base stations under some conditions. The negotiation on these conditions between the access points and the base stations can be modeled using a contract theory. Further details of the two pricing models are as follows.

1) *Utility maximization*: In the first scenario, the authors in [284] considered allocating the heterogeneous bandwidth including the cloud and WiFi bandwidth to the mobile users by using the NUM framework (see Section III-C). The utility of each mobile user is a strictly concave function of the heterogeneous bandwidth allocated to the mobile user. The method of Lagrange multipliers is used to solve the optimization problem with their interpretations as cloud bandwidth and WiFi bandwidth prices that the users are willing to pay. At each iteration, the bandwidth prices are updated by using the gradient projection method. Then, the allocated bandwidth is updated by the SDN controller. It was proved that there always exists a unique optimal solution for the allocated bandwidth and the prices. The simulation showed that the proposed approach outperforms the baseline approach in terms of the total network utility. The baseline approach did not use SDN as well as heterogeneous bandwidth. However, as stated in [257] and [258], multiple SDN controllers need to be considered to enhance the scalability of the system.

2) *Contract theory*: In the second scenario, the contract model was adopted which allows to construct several traffic-payment bundles, e.g., the amount of traffic that the access point needs to offload and the corresponding payment that the base station needs to offer. The authors in [285] used three contract theoretic models for the service trading between the base station and the access points in SDN, namely *perfect discrimination*, *linear pricing*, and *anti adverse selection*. In the *perfect discrimination*, there does not exist the information asymmetry, meaning that the idle capacities of the access points are known by the base station. In this scenario, the base station can solve its payoff optimization problems separately for each access point. The base station's payoff associated with an access point is defined as the monetary gain through

the offloaded traffic minus its payment to the access point, given the constraint that the access point's payoff is equal to or greater than zero. By taking the derivative of the objective function, an optimal payment to the access point and the amount of offloaded traffic can be obtained. The optimal payment is that the marginal valuation equals the marginal cost. On the contrary, in the *linear pricing*, the base station does not know access points' idle capacities, but it has knowledge of the probability that the access point has a certain idle capacity. Therefore, the base station only specifies a unit of traffic per payment for the offloading process. In particular, the base station formulates its payoff optimization problem by calculating the expected payment requested by the access points through the probability functions. Then, the optimal amount of traffic per payment to the access points is obtained by taking the first derivative of the objective function.

Obviously, the payoff of the base station in the *perfect discrimination* is higher than that in the *linear pricing*. However, in the *perfect discrimination*, there is no compatible incentive for the access points since their payoffs are zero. To obtain an incentive compatible contract, the *anti adverse selection* is used. The *anti adverse selection* is similar to the *perfect discrimination*. However, the constraint of the incentive compatibility for access points is added into the optimization problem of the *perfect discrimination* (i.e., the base station's payoff maximization). This problem was then solved by using the Lagrange multiplier method to determine the optimal contract, i.e., the traffic-payment, for each access point. As shown in the simulation results, the *linear pricing* gives the access points the highest payoff, followed by the *anti adverse selection*, and then the *perfect discrimination*. On the contrary, the base station gets the maximum payoff in the *perfect discrimination* since it has full knowledge of access points' idle capacities.

D. Stackelberg Game for Bandwidth Allocation in SDN-Enabled Home Networks

SDN can be combined with the existing home networks as shown in Fig. 24 to maintain users' Quality of Experience (QoE) and guarantee the QoE [286]. In this setting, home networks are connected to the service providers via a digital subscribe line or broadband cable link, and users request content services from the service provider through the subscribe lines.

Typically, service providers rent cloud bandwidth from cloud providers to deliver their contents to users. The authors in [287] investigated maximizing payoffs for the service provider and users by adopting the Stackelberg game. The service provider, i.e., the leader, charges the users, i.e., followers, for their requesting services via SDN according to the time-dependent usage-based pricing strategy [288]. The service provider also gives a reimbursement to the user, which is proportional to the amount of traffic that the user shares with other neighboring users. Thus, the payoff of the service provider is the difference between the total service price from the users and the total reimbursement. The user's payoff is the utility function of the allocated resources minus the

charge that it pays the service provider. Given the service provider's pricing and reimbursement strategy, the users aim to maximize their payoffs by choosing traffic consumption and the amount of sharing bandwidth. The optimization problems of users are solved based on the Lagrange multiplier with the subgradient projection method. Based on these optimal solutions, the service provider determines the pricing and reimbursement strategy so as to maximize its payoff function. The numerical results showed that the payoffs of both the user and the service provider from the proposed approach improve 400% compared with that in the best-effort home network with usage-based pricing [288]. However, the impact of limited backhaul capacity on the payoff functions needs to be considered in the future work.

We close this section with an approach which addresses the price-based bandwidth allocation for control applications in SDN as proposed in [289] and [290]. The aim is to maximize the rate of control applications while guaranteeing the fairness of allocation among them. The fairness criterion means that the rate of a control application is proportional to the price of bandwidth paid by the control application. The optimization problem is to maximize the sum of the rate of each control application multiplied with the corresponding price given the limited capacity. The optimization problem is a strictly convex function of the allocated rate. Therefore, there exists a unique optimal solution for the rate for each control application. However, the flow table, an essential network resource in SDN, needs to be considered in the future work.

Summary: This section discusses the applications of economic and pricing models in cloud-based SDN. The existing approaches which address the resource management in the network are summarized in Table XIV. In general, the number of existing approaches is relatively small, and most of them investigated the bandwidth allocation. More studies need to be done, for example, for the mobile data offloading.

IX. REVIEW SUMMARY, OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

A number of approaches reviewed in this survey show the importance of economic aspects not only for business development, but also for system design and optimization. Evidently, economic and pricing models have addressed various issues in cloud networking models in which traditional algorithms becomes less effective or cannot be applied. Apart from the existing approaches, there are still challenges, open issues, and new research directions as discussed in the following.

1) *False-name bidding in double auction:* In the reviewed approaches based on double auction [190], [207], [209], [210], [229], multiple cloud resources are sold simultaneously. As such, one bidder can submit multiple bids using different identities so as to gain an additional profit. The bidding process which is implemented under fictitious identities is called false-name bidding [291], and the double auction with the false-name bidding is no longer dominant-strategy incentive compatible. Therefore, robust mechanisms against the false-name bidding, e.g., the Threshold Price Double (TPD) auction protocol [292], for cloud bandwidth reservation need to be investigated.

TABLE XIV
APPLICATIONS OF ECONOMIC AND PRICING MODELS FOR RESOURCE MANAGEMENT IN CLOUD-BASED SDWN

	Ref.	Pricing model	Market structure			Mechanism	Objective	Solution
			Seller	Buyer	Item			
Cloud-based SDWN	[55]	Cost model	Mobile network operator	Users	Networking resources	Buyer evaluates the costs of the LTE network when SDN is enabled. Two types of cost are introduced, i.e. CAPEX (fixed costs) and OPEX (variable costs)	Seller's profit maximization	Cost optimization
	[56]	Cost model	Mobile network operator	Users	Networking resources	Same as [55], but three scenarios are considered when evaluating the operation cost of the LTE network	Seller's profit maximization	Cost optimization
	[86]	Bargaining game	Selling SPs	Buying SPs	CPU, memory, and bandwidth	The objective is to maximize the total increased utility of all buyers and sellers. The deal price is determined based on the Cardano's formula, and the total allocated resource is obtained using the first derivative	Increased utility maximization for both buyers and sellers	Nash bargaining solution
	[284]	Utility maximization	SP	Users	Bandwidth	The method of Lagrange multipliers is adopted. At each iteration, the bandwidth prices are updated by using the gradient projection method, and then the allocated bandwidth for buyers is updated by the seller	Total network utility maximization	Optimal solution
	[285]	Contract theory	Access points	Base station	Mobile data offloading service	Three contract theoretic models for the service trading are used, that are <i>perfect discrimination</i> , <i>linear pricing</i> , and <i>anti adverse selection</i> . In particular for the <i>anti adverse selection</i> , the seller determines the optimal prices and the amounts of bandwidth using the Lagrange multiplier method	Payoff maximization, incentive compatibility, and individual rationality	Optimal solution
	[287]	Stackelberg game	SP	Users	Bandwidth	Seller sets the bandwidth price based on the time-dependent usage-based pricing strategy. Then, buyers choose the amount of bandwidth based on the Lagrange multiplier method. The seller finally determines the bandwidth price	Payoff maximization for sellers and buyers	Stackelberg equilibrium

TABLE XV
A SUMMARY OF ADVANTAGES AND DISADVANTAGES OF MAJOR APPROACHES FOR THE RESOURCE MANAGEMENT IN CLOUD-BASED SDWN.

Major approaches	Advantages	Disadvantages
[86]	<ul style="list-style-type: none"> Achieve flexible resource management and demand-driven resource distribution 	<ul style="list-style-type: none"> Be not appropriate in the realistic scenarios
[284]	<ul style="list-style-type: none"> Manage resource in a centralized and holistic manner Enable reliable functional verification 	<ul style="list-style-type: none"> Be unscalable Do not support heterogeneous resource allocation Support only one SDN controller
[285]	<ul style="list-style-type: none"> Overcome the information asymmetry Support multiple traffic-payment bundles 	<ul style="list-style-type: none"> Support only one SDN controller
[287]	<ul style="list-style-type: none"> Adapt to both real-time and non-real-time service requests Be resilient to demand fluctuations 	<ul style="list-style-type: none"> Support only one service provider

2) *Collusion in auction*: Apart from the false-name bidding cheating, bidders in the reviewed approaches based on auction, i.e., the VCG auction [124], [224], the combinatorial auction [192], and the double auction [190], [207], [209], [210], [229], may collude with each other through coordinating their bids. This suppresses the competition for cloud resource, thus reducing the price that the bidders must pay for the cloud resource. However, the collusion behavior will degrade the efficiency of the resource allocation as well as the cloud provider's revenue. Collusion-resistant mechanisms thus need to be applied. Pricing strategies can be used to provide the bidders incentives not to perform the collusion behavior.

3) *Resource demand forecasting methods*: As discussed in [158] (see Section IV-A8), resource demand fluctuation of cloud tenants impacts the availability of resources, the pricing policy as well as the profit of cloud providers. Therefore, it is important for the cloud provider to predict the workload fluctuation. However, there was only one technique intro-

duced for the demand forecast, i.e., the Markov chain model. Therefore, more advanced techniques need to be adopted to improve the performance of resource demand prediction in cloud networking environment. Some candidate methods are fuzzy logic [293], neural networks [294], machine learning [295], and a joint statistical learning and optimal decision-making [296].

4) *Pricing models for multipath routing in cloud data center networking*: Data centers consist of several hundreds and thousands of servers [297]. The servers are connected with each other through multi-rooted hierarchical topologies, e.g., the fat tree topology [298], which provide multipath between each pair of servers. A simple example is shown in Fig. 25 in which there are two different paths, i.e., a1 and a2, between servers 1 and 2. Efficient routing strategies are thus necessary. Traditional routing algorithms such as distance vector [299] and link state [300] do not support the multipath routing efficiently because they make routing decisions only based on

packet destinations. This means that all packets to the same destination are routed through the same path which may result in congestion. By taking into account the multipath feature, pricing approaches such as auction are efficient solutions for the path selection to achieve a highly load-balanced network and minimize network latency. For example, the reverse auction can be used at the switch A which acts as a buyer to select path a1 or a2 based on asking prices submitted from switches B and C, i.e., sellers. The asking price of switch B or C is proportional to the total congestion level of links connected to the switch and the number of hops from the switch towards the destination. The path containing the lowest asking price is selected for forwarding the packet.

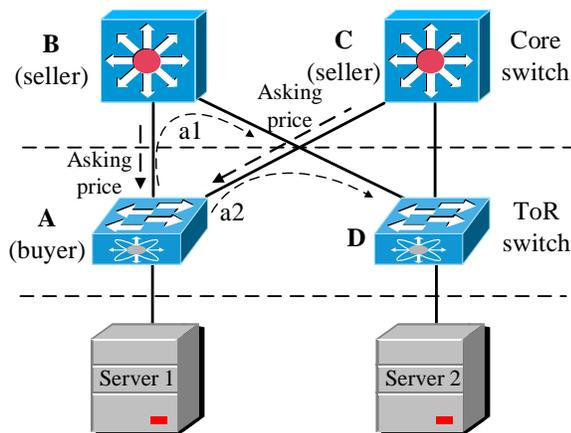


Fig. 25. Multipath routing in data center networks based on reserve auction.

5) *Pricing models in cloud robotics*: Cloud robotics is a combination of robotic technology with network and cloud computing infrastructure [301], [302]. Robots benefit from powerful computing, storage, and communication resources of data centers in the cloud. This integration enables robotic systems to be smarter, faster, and less expensive. A common cloud robotic system is shown in Fig. 26. Multiple robots are connected with data centers through wired or wireless networks to perform tasks. The proxy allocates resources in the cloud to the client robots. Since robotic applications require real-time execution, the key challenge is the low-latency response given network bandwidth constraint. Therefore, efficient bandwidth allocation among robots is of great importance. Resource management mechanisms such as auction or Stackelberg game are promising solutions since they model an interaction among client robots. For example, auctions can be adopted among client robots, i.e., buyers, which compete for the connectivity provided by a relay robot, i.e., a seller [303]. To maximize the utility of relay or client robots and cloud provider's profits, the Stackelberg game can be used [304], [305] in which the cloud provider, i.e., the leader, optimizes bandwidth price, and client robots, i.e., followers, choose their transmission rates.

6) *Economic and pricing models in Network Function Virtualization (NFV)*: Network Function Virtualization (NFV) is a concept that leverages virtualization technologies to offer a new way for designing, deploying, and managing networking services. NFV and SDN are closely related technologies which are implemented through the software running on physical

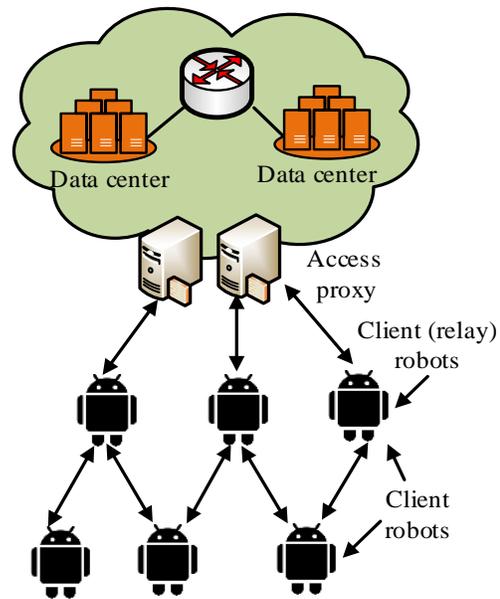


Fig. 26. An architecture of typical cloud robotic systems.

equipments. While SDN decouples the network control and forwarding functions on a physical network equipment, NFV decomposes network functions, e.g., a firewall, from the physical network equipment [306]. NFV brings several benefits such as reduction of OPEX and CAPEX due to consolidating networking appliances [307], facilitating the deployment of new services with increased agility and faster time-to-value, and achieving better system scalability according to users' demand. However, implementing NFV introduces several challenges. For example, physical resource sharing among multiple users can lead to congestion on the physical infrastructure as well as an unfair use of the resources [308], [309]. Pricing approaches such as auction or smart data pricing can be adopted to offer incentives to users to use resources efficiently.

X. CONCLUSIONS

This paper has presented a comprehensive survey of the applications of economic and pricing theories to resource management in cloud networking. Firstly, we have described a general architecture of the cloud networking including its components and corresponding services. Then, we have introduced and analyzed various pricing models with the objectives to understand the motivations of using the economic and pricing theory in cloud networking. Afterwards, we have provided detailed reviews, analyses, and comparisons of the approaches using economic and pricing theories to solve a variety of issues in specific systems of cloud networking, i.e., the cloud data center networking, mobile cloud networking, edge computing, cloud-based VoD system, and cloud-based SDWN. Finally, we have outlined open issues as well as future research directions.

REFERENCES

- [1] N. Sultan, "Cloud computing for education: A new dawn?" *International Journal of Information Management*, vol. 30, no. 2, pp. 109–116, Apr. 2010.

- [2] M. R. Rahimi, J. Ren, C. H. Liu, A. V. Vasilakos, and N. Venkatasubramanian, "Mobile cloud computing: A survey, state of art and future directions," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 133–143, Apr. 2014.
- [3] M.-H. Kuo, "Opportunities and challenges of cloud computing to improve health care services," *Journal of medical Internet research*, vol. 13, no. 3, pp. 67–89, Sep. 2011.
- [4] Z. Li, C. Chen, and K. Wang, "Cloud computing for agent-based urban transportation systems," *IEEE Intelligent Systems*, vol. 26, no. 1, pp. 73–79, Feb. 2011.
- [5] K. Chard, S. Caton, O. F. Rana, and K. Bubendorfer, "Social cloud: Cloud computing in social networks." Miami, FL, USA, Apr. 2010, pp. 99–106.
- [6] (2011) Google docs. [Online]. Available: <http://docs.google.com>
- [7] (2011) Google app engine. [Online]. Available: <http://code.google.com/appengine/>
- [8] (2011) Amazon virtual private cloud. [Online]. Available: <http://aws.amazon.com/ec2/>
- [9] (2012) Cloud network architecture description. [Online]. Available: [whhttp://www.sail-project.eu/wp-content/uploads/2012/06/D-D.1v2.0-final-public.pdf](http://www.sail-project.eu/wp-content/uploads/2012/06/D-D.1v2.0-final-public.pdf)
- [10] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: a survey," *IEEE Communications Magazine*, vol. 51, no. 11, pp. 24–31, Nov. 2013.
- [11] Q. Duan, Y. Yan, and A. V. Vasilakos, "A survey on service-oriented network virtualization toward convergence of networking and cloud computing," *IEEE Transactions on Network and Service Management*, vol. 9, no. 4, pp. 373–392, Dec. 2012.
- [12] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 27–51, Jun. 2015.
- [13] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmoly, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.
- [14] H. Farhady, H. Lee, and A. Nakao, "Software-defined networking: A survey," *Computer Networks*, vol. 81, no. 1, pp. 79–95, Apr. 2015.
- [15] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu, and A. V. Vasilakos, "Software-defined and virtualized future mobile and wireless networks: A survey," *Mobile Networks and Applications*, vol. 20, no. 1, pp. 4–18, Feb. 2015.
- [16] S. Yi, C. Li, and Q. Li, "A survey of fog computing: concepts, applications and issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*. Hangzhou, China: ACM, Jun. 2015, pp. 37–42.
- [17] H. He, K. Xu, and Y. Liu, "Internet resource pricing models, mechanisms, and methods," *Networking Science*, vol. 1, no. 1, pp. 48–66, Mar. 2012.
- [18] L. A. DaSilva, "Pricing for qos-enabled networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 3, no. 2, pp. 2–8, Nov. 2000.
- [19] C. A. Gizelis and D. D. Vergados, "A survey of pricing schemes in wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 1, pp. 126–145, Jul. 2011.
- [20] P. Murray, A. Sefidcon, R. Steinert, V. Fusenig, and J. Carapinha, "Cloud networking: an infrastructure service architecture for the wide area," in *Future Network & Mobile Summit (FutureNetw)*, Berlin, Germany, Jul. 2012, pp. 1–8.
- [21] H. T. Mouftah, *Communication Infrastructures for Cloud Computing*. IGI Global, 2013.
- [22] N. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Computer Networks*, vol. 54, no. 5, pp. 862–876, Apr. 2010.
- [23] X. Xiang, C. Lin, F. Chen, and X. Chen, "Greening geo-distributed data centers by joint optimization of request routing and virtual machine scheduling," in *Proceedings of the IEEE/ACM 7th International Conference on Utility and Cloud Computing*, London, United Kingdom, Dec. 2014, pp. 1–10.
- [24] N. Bitar, S. Gringeri, and T. J. Xia, "Technologies and protocols for data center and cloud networking," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 24–31, Sep. 2013.
- [25] A. Levin and P. Massonet, "Enabling federated cloud networking," in *Proceedings of the 8th ACM International Systems and Storage Conference*. Haifa, Israel: ACM, May 2015, pp. 23–23.
- [26] A. Jamakovic, T. M. Bohnert, and G. Karagiannis, *Mobile Cloud Networking: Mobile Network, Compute, and Storage as One Service On-Demand*. Springer, 2013.
- [27] G. Karagiannis, A. Jamakovic, A. Edmonds, C. Parada, T. Metsch, D. Pichon, M. Corici, S. Ruffino, A. Gomes, P. S. Crosta *et al.*, "Mobile cloud networking: Virtualisation of cellular networks," in *21st International Conference on Telecommunications (ICT)*, Lisbon, Portugal, May 2014, pp. 410–415.
- [28] G. Lewis, S. Echeverría, S. Simanta, B. Bradshaw, and J. Root, "Tactical cloudlets: Moving cloud computing to the edge," in *IEEE Military Communications Conference*, Baltimore, MD, USA, Oct. 2014, pp. 1440–1446.
- [29] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, Warsaw, Poland, Sep. 2014, pp. 1–8.
- [30] B. Ahlgren, P. A. Aranda, P. Chemouil, S. Oueslati, L. M. Correia, H. Karl, M. Söllner, and A. Welin, "Content, connectivity, and cloud: ingredients for the network of the future," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 62–70, Jun. 2011.
- [31] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. of the Sixth International Conference on Advances in Future Internet*, Lisbon, Portugal, Nov. 2014, pp. 1–7.
- [32] J. E. Smith and R. Nair, "The architecture of virtual machines," *Computer*, vol. 38, no. 5, pp. 32–38, May 2005.
- [33] Y. Yuan, C.-r. Wang, and C. Wang, "A game based approach for sharing the data center network," in *Advances in Neural Networks-ISNN 2012*. Springer, 2012, pp. 641–649.
- [34] G. Carella, A. Edmonds, F. Dudouet, M. Corici, B. Sousa, and Z. Yousaf, "Mobile cloud networking: From cloud, through nfv and beyond," in *IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN)*, San Francisco, CA, USA, Nov. 2015, pp. 7–8.
- [35] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, Sep. 2014.
- [36] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Communications*, vol. 22, no. 2, pp. 152–160, Apr. 2015.
- [37] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. Di Girolamo, and G. Giuliani, "Ran as a service: Challenges of designing a flexible ran architecture in a cloud-based heterogeneous mobile network," in *Future Network and Mobile Summit (FutureNetworkSummit)*, Lisbon, Portugal, Jul. 2013, pp. 1–8.
- [38] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, "Edge-centric computing: Vision and challenges," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 5, pp. 37–42, Oct. 2015.
- [39] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *the Proceedings of the 10th IEEE International Conference on Intelligent Systems and Control (ISCO 2016)*, Coimbatore, India, Jan. 2016, pp. 1–8.
- [40] B. Raghavan, M. Casado, T. Koponen, S. Ratnasamy, A. Ghodsi, and S. Shenker, "Software-defined internet architecture: decoupling architecture from infrastructure," in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. Redmond, WA, USA: ACM, Oct. 2012, pp. 43–48.
- [41] J. Pan, S. Paul, and R. Jain, "A survey of the research on future internet architectures," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 26–36, Jul. 2011.
- [42] S. Zerrik, R. Atay, M. Bakhouya, J. Gaber *et al.*, "Towards a decentralized and adaptive software-defined networking architecture," in *International Conference on Next Generation Networks and Services (NGNS)*, Casablanca, Morocco, May 2014, pp. 326–329.
- [43] S. Costanzo, L. Galluccio, G. Morabito, and S. Palazzo, "Software defined wireless network (sdwn): An evolvable architecture for w-pans," in *IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, Torino, Italy, Sep. 2015, pp. 23–28.
- [44] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1617–1634, Feb. 2014.
- [45] A. K. Nayak, A. Reimers, N. Feamster, and R. Clark, "Resonance: dynamic access control for enterprise networks," in *Proceedings of the 1st ACM workshop on Research on enterprise networking*. Barcelona, Spain: ACM, Aug. 2009, pp. 11–18.
- [46] N. Handigol, S. Seetharaman, M. Flajslik, N. McKeown, and R. Johari, "Plug-n-serve: Load-balancing web traffic using openflow," *ACM Sigcomm Demo*, vol. 4, no. 5, pp. 6–7, Aug. 2009.
- [47] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu *et al.*, "B4: Experience with a globally-deployed software defined wan," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 3–14, Aug. 2013.

- [48] M. G. Rabbani, R. P. Esteves, M. Podlesny, G. Simon, L. Z. Granville, and R. Boutaba, "On tackling virtual data center embedding problem," in *IFIP/IEEE International Symposium on Integrated Network Management*, Ghent, Belgium, May 2013, pp. 177–184.
- [49] L. Suresh, J. Schulz-Zander, R. Merz, A. Feldmann, and T. Vazao, "Towards programmable enterprise wlangs with odin," in *Proceedings of the first workshop on Hot topics in software defined networks*. Helsinki, Finland: ACM, Aug. 2012, pp. 115–120.
- [50] L. Liu, T. Tsuritani, I. Morita, H. Guo, and J. Wu, "Openflow-based wavelength path control in transparent optical networks: a proof-of-concept demonstration," in *European Conference and Exposition on Optical Communications*. Geneva, Switzerland: Optical Society of America, Sep. 2011, pp. 1–5.
- [51] K. L. Calvert, W. K. Edwards, N. Feamster, R. E. Grinter, Y. Deng, and X. Zhou, "Instrumenting home networks," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, pp. 84–89, Jan. 2011.
- [52] C. Courcoubetis and R. Weber, *Cost-Based Pricing*. John Wiley & Sons, Ltd, 2003, pp. 161–194. [Online]. Available: <http://dx.doi.org/10.1002/0470867175.ch7>
- [53] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68–73, Jan. 2008.
- [54] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: Automated data placement for geo-distributed cloud services," in *NSDI*, San Jose, California, USA, Sep. 2010, pp. 17–32.
- [55] N. Zhang and H. Hammainen, "Cost efficiency of sdn in lte-based mobile networks: Case finland," in *International Conference and Workshops on Networked Systems (NetSys)*, Cottbus, Germany, Mar. 2015, pp. 1–5.
- [56] T. M. Knoll, "Life-cycle cost modelling for nvf/sdn based mobile networks," in *Conference of Telecommunication, Media and Internet Techno-Economics (CTTE)*, Munich, Germany, Nov. 2015, pp. 1–8.
- [57] R. S. Pindyck and D. L. Rubinfeld, "Microeconomics, 6," *Auflage, New Jersey*, pp. 613–640, 2005.
- [58] T. T. Nagle, J. E. Hogan, and J. Zale, *The Strategy and Tactics of Pricing: A Guide to Growing More Profitably*. Pearson/Prentice Hall Upper Saddle River, NJ, 2006.
- [59] D. M. Divakaran, M. Gurusamy, and M. Sellamuthu, "Bandwidth allocation with differential pricing for flexible demands in data center networks," *Computer Networks*, vol. 73, no. 1, pp. 84–97, Nov. 2014.
- [60] R. Korrapati, *Validated Management Practices*, ser. A.H.W. Sameer series. Diamond Pocket Books, 2014. [Online]. Available: <https://books.google.com.sg/books?id=z4psBQAAQBAJ>
- [61] L. Tsai and W. Liao, *Virtualized Cloud Data Center Networks: Issues in Resource Management*, ser. Springerbriefs in electrical and computer engineering. Cham, Switzerland: Springer, 2016.
- [62] K. Tsakalozos, H. Kllapi, E. Sitaridi, M. Roussopoulos, D. Pappas, and A. Delis, "Flexible use of cloud resources through profit maximization and price discrimination," in *IEEE International Conference on Data Engineering (ICDE)*, Hannover, Germany, Apr. 2011, pp. 75–86.
- [63] S. Rebai, M. Hadji, and D. Zeghlache, "Improving profit through cloud federation," in *12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, USA, Jan. 2015, pp. 732–739.
- [64] W. G. Shepherd, "Ramsey pricing: Its uses and limits," *Utilities Policy*, vol. 2, no. 4, pp. 296–298, Oct. 1992.
- [65] U. S. C. B. Office, *Paying for Highways, Airways, and Waterways: How Can Users be Charged?* Congressional Budget Office, 1992.
- [66] B. Wanis, N. Samaan, and A. Karmouch, "Efficient modeling and demand allocation for differentiated cloud virtual-network as-a service offerings," *IEEE Transactions on Cloud Computing*, to appear.
- [67] R. Gibbons, *A Primer In Game Theory*. Harvester Wheatsheaf, 1992.
- [68] H.-Y. Shi, W.-L. Wang, N.-M. Kwok, and S.-Y. Chen, "Game theory for wireless sensor networks: a survey," *Sensors*, vol. 12, no. 7, pp. 9055–9097, Jul. 2012.
- [69] T. AlSkaif, M. G. Zapata, and B. Bellalta, "Game theory for energy efficiency in wireless sensor networks: Latest trends," *Journal of Network and Computer Applications*, vol. 54, no. 1, pp. 33–61, Aug. 2015.
- [70] J. W. Friedman, *A non-cooperative equilibrium for supergames*. Cambridge Univ Pr, 1988.
- [71] J. Chakareski, "Cost and profit driven cloud-p2p interaction," *Peer-to-Peer Networking and Applications*, vol. 8, no. 2, pp. 244–259, Mar. 2015.
- [72] Z. Guan and T. Melodia, "The value of cooperation: Minimizing user costs in multi-broker mobile cloud computing networks," *IEEE Transactions on Cloud Computing*, to appear.
- [73] R. Pal and P. Hui, "Economic models for cloud service markets: Pricing and capacity planning," *Theoretical Computer Science*, vol. 496, no. 1, pp. 113–124, Jul. 2013.
- [74] R. Amir and I. Grilo, "Stackelberg versus cournot equilibrium," *Games and Economic Behavior*, vol. 26, no. 1, pp. 1–21, May 1999.
- [75] S. Kim, *Game theory applications in network design*. IGI Global, 2014.
- [76] Z. Han, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2012.
- [77] G. Leitmann, *Multicriteria Decision Making and Differential Games*. Springer, 2013.
- [78] Y. Lin and H. Shen, "Autotune: game-based adaptive bitrate streaming in p2p-assisted cloud-based vod systems," in *IEEE International Conference on Peer-to-Peer Computing (P2P)*, Cambridge, MA, USA, Sep. 2015, pp. 1–10.
- [79] A. Al Daoud, S. Agarwal, and T. Alpcan, "Brief announcement: Cloud computing games: Pricing services of large data centers," in *International Symposium on Distributed Computing*. Springer, Jan. 2009, pp. 309–310.
- [80] V. Di Valerio, V. Cardellini, and F. L. Presti, "Optimal pricing and service provisioning strategies in cloud systems: a stackelberg game approach," in *IEEE Sixth International Conference on Cloud Computing*, Santa Clara Marriott, CA, USA, Jul. 2013, pp. 115–122.
- [81] X. Lv, R. Zhang, and J. Yue, "Competition and cooperation between participants of the internet of things industry value chain," *Advances in Information Sciences & Service Sciences*, vol. 4, no. 11, pp. 406–412, Jun. 2012.
- [82] A. Danak, A. R. Kian, and B. Moshiri, "Inner supervision in multi-sensor data fusion using the concepts of stackelberg games," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Heidelberg, Germany, Sep. 2006, pp. 65–70.
- [83] A. Muthoo, *Bargaining Theory With Applications*. Cambridge University Press, 1999.
- [84] K. Chatterjee and W. Samuelson, "Bargaining under incomplete information," *Operations Research*, vol. 31, no. 5, pp. 835–851, Aug. 1983.
- [85] H. Xu and B. Li, "A general and practical datacenter selection framework for cloud services," in *IEEE 5th International Conference on Cloud Computing (CLOUD)*, Honolulu, Hawaii, USA, Jun. 2012, pp. 9–16.
- [86] J. Ding, R. Yu, Y. Zhang, S. Gjessing, and D. H. Tsang, "Service provider competition and cooperation in cloud-based software defined wireless networks," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 134–140, Nov. 2015.
- [87] P. Samimi and A. Patel, "Review of pricing models for grid & cloud computing," in *IEEE Symposium on Computers & Informatics (ISCI)*, Kuala Lumpur, Malaysia, Mar. 2011, pp. 634–639.
- [88] G. N. Iyer and B. Veeravalli, "On the resource allocation and pricing strategies in compute clouds using bargaining approaches," in *IEEE International Conference on Networks*, Singapore, Dec. 2011, pp. 147–152.
- [89] R. P. McAfee and J. McMillan, "Auctions and bidding," *Journal of economic literature*, vol. 25, no. 2, pp. 699–738, Jun. 1987.
- [90] P. Klemperer, *Auctions: Theory and Practice*, ser. Princeton paperbacks. Princeton University Press, 2004. [Online]. Available: <https://books.google.com.vn/books?id=-FJjQgAACAAJ>
- [91] K. Chui and R. Zwick. (1999) Auction on the internet-a preliminary study. [Online]. Available: <http://repository.ust.hk/ir/Record/1783.1-1035>
- [92] Y. Zhang, D. Niyato, and P. Wang, "An auction mechanism for resource allocation in mobile cloud computing systems," in *Wireless Algorithms, Systems, and Applications*. Springer, 2013, pp. 76–87.
- [93] K. Vijay, *Auction Theory*. Academic Press, 2009.
- [94] V. Rodriguez and F. Jondral, "Simple adaptively-prioritised spatially-reusable medium access control through the dutch auction: Qualitative analysis, issues, challenges," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux*, Benelux, Nov. 2007, pp. 1–5.
- [95] D. Lucking-Reiley, "Vickrey auctions in practice: From nineteenth-century philately to twenty-first-century e-commerce," *The Journal of Economic Perspectives*, vol. 14, no. 3, pp. 183–192, Feb. 2000.
- [96] T. W. Sandholm, "Limitations of the vickrey auction in computational multiagent systems," in *Proceedings of the Second International Conference on Multiagent Systems (ICMAS-96)*, Kyoto, Japan, Dec. 1996, pp. 299–306.
- [97] P. Klemperer, "What really matters in auction design," *The Journal of Economic Perspectives*, vol. 16, no. 1, pp. 169–189, Sep. 2002.

- [98] L. M. Ausubel, P. Milgrom *et al.*, “The lovely but lonely vickrey auction,” *Combinatorial auctions*, vol. 17, pp. 22–26, Aug. 2006.
- [99] D. Friedman and J. Rust, *The Double Auction Market: Institutions, Theories, and Evidence*. Westview Press, 1993.
- [100] T. Mullen and M. P. Wellman, “Market-based negotiation for digital library services,” in *Second USENIX Workshop on Electronic Commerce*, vol. 13, Oakland, California, USA, Nov. 1996, pp. 259–269.
- [101] V. Krishna, *Auction theory*. Academic press, 2002.
- [102] P. C. Cramton, Y. Shoham, R. Steinberg *et al.*, *Combinatorial Auctions*. MIT press Cambridge, 2006, vol. 475.
- [103] F.-S. Hsieh, “Combinatorial reverse auction based on revelation of lagrangian multipliers,” *Decision Support Systems*, vol. 48, no. 2, pp. 323–330, Jan. 2010.
- [104] M. Pan, F. Chen, X. Yin, and Y. Fang, “Fair profit allocation in the spectrum auction using the shapley value,” in *IEEE GLOBECOM*, Honolulu, Hawaii, USA, Dec. 2009, pp. 1–6.
- [105] A. E. Roth, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988.
- [106] R. Kaewpuang, D. Niyato, P. Wang, and E. Hossain, “A framework for cooperative resource management in mobile cloud computing,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 2685–2700, Dec. 2013.
- [107] L. Shapley, “A value for n-person games I,” *Contributions to the Theory of Games (AM-28)*, vol. 2, pp. 307–399, Aug. 2016.
- [108] E. Winter, “The shapley value,” *Handbook of game theory with economic applications*, vol. 3, no. 1, pp. 2025–2054, Dec. 2002.
- [109] R. van den Brink, Y. Funaki, and Y. Ju, “Reconciling marginalism with egalitarianism: consistency, monotonicity, and implementation of egalitarian shapley values,” *Social Choice and Welfare*, vol. 40, no. 3, pp. 693–714, Mar. 2013.
- [110] D. Niyato, K. Zhu, and P. Wang, “Cooperative virtual machine management for multi-organization cloud computing environment,” in *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, Paris, France, May 2011, pp. 528–537.
- [111] R. T. Ma, D. M. Chiu, J. Lui, V. Misra, and D. Rubenstein, “Internet economics: The use of shapley value for isp settlement,” *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 3, pp. 775–787, May 2010.
- [112] K. Suzuki, K. Kobayashi, and H. Morita, “Efficient sealed-bid auction using hash chain,” in *Information Security and Cryptology ICISC 2000*. Springer, 2001, pp. 183–191.
- [113] J. Hartline. (2002) Dynamic posted price mechanisms. [Online]. Available: <http://www.ece.northwestern.edu/~hartline/papers/posted-price.pdf>
- [114] C.-C. Wu, Y.-F. Liu, Y.-J. Chen, and C.-J. Wang, “Consumer responses to price discrimination: Discriminating bases, inequality status, and information disclosure timing influences,” *Journal of Business Research*, vol. 65, no. 1, pp. 106–116, Jan. 2012.
- [115] A. Badanidiyuru, R. Kleinberg, and Y. Singer, “Learning on a budget: posted price mechanisms for online procurement,” in *Proceedings of the 13th ACM Conference on Electronic Commerce*, Valencia, Spain, Jun. 2012, pp. 128–145.
- [116] V. Nallur and R. Bahsoon, “A decentralized self-adaptation mechanism for service-based applications in the cloud,” *IEEE Transactions on Software Engineering*, vol. 39, no. 5, pp. 591–612, May 2013.
- [117] J. Sun and H. Ma, “Collection-behavior based multi-parameter posted pricing mechanism for crowd sensing,” in *IEEE ICC*, Sydney, Australia, Jun. 2014, pp. 227–232.
- [118] J. Sun. (2013) Behavior-based online incentive mechanism for crowd sensing with budget constraints. [Online]. Available: <https://arxiv.org/pdf/1310.5485.pdf>
- [119] A. Mas-Colell, M. D. Whinston, J. R. Green *et al.*, *Microeconomic Theory*. Oxford university press New York, 1995, vol. 1.
- [120] D. Bertsekas and A. Nedic, *Convex Analysis and Optimization*. Athena Scientific, 2003.
- [121] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, “Non-convex optimization and rate control for multi-class services in the internet,” *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 827–840, Aug. 2005.
- [122] D. Abts and B. Felderman, “A guided tour through data-center networking,” *Queue*, vol. 10, no. 5, pp. 10–24, May 2012.
- [123] J. Chase and D. Niyato, “Joint optimization of resource provisioning in cloud computing,” *IEEE Transactions on Services Computing*, to appear.
- [124] Y. Gui, Z. Zheng, F. Wu, X. Gao, and G. Chen, “Soar: Strategy-proof auction mechanisms for distributed cloud bandwidth reservation,” in *IEEE International Conference on Communication Systems (ICCS)*, Macau, Nov. 2014, pp. 162–166.
- [125] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT press, 1994.
- [126] A. Mu’Alem and N. Nisan, “Truthful approximation mechanisms for restricted combinatorial auctions,” *Games and Economic Behavior*, vol. 64, no. 2, pp. 612–631, Nov. 2008.
- [127] D. Lehmann, L. I. O’callaghan, and Y. Shoham, “Truth revelation in approximately efficient combinatorial auctions,” *Journal of the ACM (JACM)*, vol. 49, no. 5, pp. 577–602, Sep. 2002.
- [128] R. Lavi and C. Swamy, “Truthful and near-optimal mechanism design via linear programming,” *Journal of the ACM (JACM)*, vol. 58, no. 6, p. 25, Dec. 2011.
- [129] W. Shi, C. Wu, and Z. Li, “A shapley-value mechanism for bandwidth on demand between datacenters,” *IEEE Transactions on Cloud Computing*, to appear.
- [130] W. K. Tan, D. M. Divakaran, and M. Gurusamy, “Uniform price auction for allocation of dynamic cloud bandwidth,” in *IEEE ICC*, Sydney, Australia, Jun. 2014, pp. 2944–2949.
- [131] C. Gittings, *The Advertising Handbook*. Routledge, New York, 2002.
- [132] J. Guo, F. Liu, D. Zeng, J. Lui, and H. Jin, “A cooperative game based allocation for sharing data center networks,” in *IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 2139–2147.
- [133] J. Guo, F. Liu, H. Tang, Y. Lian, H. Jin, and J. Lui, “Faloc: Fair network bandwidth allocation in iaas datacenters via a bargaining game approach,” in *21st IEEE International Conference on Network Protocols (ICNP)*, Gottingen, Germany, Oct. 2013, pp. 1–10.
- [134] J. Guo, F. Liu, J. Lui, and H.-J. Jin, “Fair network bandwidth allocation in iaas datacenters via a cooperative game approach,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 873–886, Jan. 2016.
- [135] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [136] F. Kelly, “Charging and rate control for elastic traffic,” *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, Jan. 1997.
- [137] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, “A game theoretic framework for bandwidth allocation and pricing in broadband networks,” *IEEE/ACM Transactions on Networking (TON)*, vol. 8, no. 5, pp. 667–678, Oct. 2000.
- [138] W. Li, D. Guo, K. Li, H. Qi, and J. Zhang, “idaas: Inter-datacenter network as a service,” *IEEE Transactions on Parallel and Distributed Systems*, to appear.
- [139] Y. Feng, B. Li, and B. Li, “Bargaining towards maximized resource utilization in video streaming datacenters,” in *IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1134–1142.
- [140] N. Nikaëin, 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 *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture*. Paris, France: ACM, Sep. 2015, pp. 8–13.
- [141] C. Wang, Y. Yuan, and C. Wan, “Lease data center in the light of network resources: An economic model,” in *Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, Harbin, China, Sep. 2014, pp. 606–610.
- [142] S. Huck, W. Muller, and H.-T. Normann, “Stackelberg beats cournot-on collusion and efficiency in experimental markets,” *The Economic Journal*, vol. 111, no. 474, pp. 749–765, Oct. 2001.
- [143] D. M. Divakaran and M. Gurusamy, “Probabilistic-bandwidth guarantees with pricing in data-center networks,” in *IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 3716–3720.
- [144] N. Kamiyama and V. O. Li, “An efficient deterministic bandwidth allocation method in interactive video-on-demand systems,” in *IEEE GLOBECOM*, Sydney, Australia, Nov. 1998, pp. 664–671.
- [145] D. M. Divakaran and M. Gurusamy, “Towards flexible guarantees in clouds: Adaptive bandwidth allocation and pricing,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 6, pp. 1754–1764, May 2015.
- [146] H. Shen and Z. Li, “New bandwidth sharing and pricing policies to achieve a win-win situation for cloud provider and tenants,” in *IEEE INFOCOM*, Toronto, Canada, May 2014, pp. 835–843.
- [147] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, and I. Stoica, “Faircloud: sharing the network in cloud computing,” in *Proceedings on Applications, technologies, architectures, and protocols for computer communication*. Helsinki, Finland: ACM, Aug. 2012, pp. 187–198.

- [148] Y. Zhan, D. Xu, and Y. Ou, "Distributednet: A reasonable pricing and flexible network architecture for datacenter," in *IEEE ICC*, Sydney, Australia, Jun. 2014, pp. 3999–4004.
- [149] Y. Zhan, D. Xu, H. Yang, M. Tang, S. Peng, and D. Simeonidou, "Adaptive purchase option for multi-tenant data center," in *IEEE ICC*, London, United Kingdom, Jun. 2015, pp. 358–363.
- [150] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "The price is right: towards location-independent costs in datacenters," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*. New York, NY, USA: ACM, Nov. 2011, pp. 23–28.
- [151] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards predictable datacenter networks," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. New York, NY, USA: ACM, Apr. 2011, pp. 242–253.
- [152] D. Stefani Marcon, R. Ruas Oliveira, M. Cardoso Neves, L. Saete Buriol, L. P. Gaspar, and M. Pilla Barcellos, "Trust-based grouping for cloud datacenters: improving security in shared infrastructures," in *IFIP Networking Conference*, Brooklyn, NY, USA, May 2013, pp. 1–9.
- [153] W. J. Baumol and D. F. Bradford, "Optimal departures from marginal cost pricing," *The American Economic Review*, vol. 60, no. 3, pp. 265–283, Jul. 1970.
- [154] T. H. Oum and M. W. Tretheway, "Ramsey pricing in the presence of externality costs," *Journal of Transport Economics and Policy*, vol. 22, no. 3, pp. 307–317, Sep. 1988.
- [155] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data center tcp (dctcp)," *ACM SIGCOMM computer communication review*, vol. 41, no. 4, pp. 63–74, Oct. 2011.
- [156] M. Mihailescu and Y. M. Teo, "Dynamic resource pricing on federated clouds," in *IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, Melbourne, Australia, May 2010, pp. 513–517.
- [157] K. Li, J. Wu, and A. Blaisse, "Elasticity-aware virtual machine placement for cloud datacenters," in *IEEE International Conference on Cloud Networking (CloudNet)*, San Francisco, USA, Nov. 2013, pp. 99–107.
- [158] B. Wanis, N. Samaan, and A. Karmouch, "Modeling and pricing cloud service elasticity for geographically distributed applications," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Ottawa, Canada, May 2015, pp. 559–565.
- [159] S. Pacheco-Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr, and S. Dawson, "Markovian workload characterization for qos prediction in the cloud," in *IEEE International Conference on Cloud Computing (CLOUD)*, Washington, DC, USA, Jul. 2011, pp. 147–154.
- [160] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific Belmont, MA, 1995.
- [161] B. Wong and E. G. Sirer, "Closestnode. com: an open access, scalable, shared geocast service for distributed systems," *ACM SIGOPS Operating Systems Review*, vol. 40, no. 1, pp. 62–64, Jan. 2006.
- [162] K. H. Prasad, T. Faruque, V. L. Subramaniam, M. Mohania, G. Venkatchalialah *et al.*, "Resource allocation and sla determination for large data processing services over cloud," in *IEEE International Conference on Services Computing (SCC)*, Miami, FL, USA, Jul. 2010, pp. 522–529.
- [163] M. Avriel, *Nonlinear Programming: Analysis and Methods*. Courier Corporation, 2003.
- [164] Q. Zhang, Q. Zhu, M. F. Zhani, and R. Boutaba, "Dynamic service placement in geographically distributed clouds," in *IEEE International Conference on Distributed Computing Systems (ICDCS)*, Macau, China, Jun. 2012, pp. 526–535.
- [165] Q. Zhang, Q. Zhu, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic service placement in geographically distributed clouds," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 762–772, Dec. 2013.
- [166] F. Wu, Q. Wu, and Y. Tan, "Workflow scheduling in cloud: a survey," *The Journal of Supercomputing*, vol. 71, no. 9, pp. 3373–3418, Sep. 2015.
- [167] A. Gupta, U. Mandai, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Cost-efficient live vm migration based on varying electricity cost in optical cloud networks," in *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, New Delhi, India, Dec. 2014, pp. 1–3.
- [168] A. Gupta, U. Mandal, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Cost-efficient live vm migration based on varying electricity cost in optical cloud networks," *Photonic Network Communications*, vol. 30, no. 3, pp. 376–386, Dec. 2015.
- [169] B. Kantarci and H. Mouftah, "Inter-data center network dimensioning under time-of-use pricing," *IEEE Transactions on Cloud Computing*, vol. 3, no. 99, pp. 1–14, Nov. 2014.
- [170] J. Blythe, S. Jain, E. Deelman, Y. Gil, K. Vahi, A. Mandal, and K. Kennedy, "Task scheduling strategies for workflow-based applications in grids," in *IEEE International Symposium on Cluster Computing and the Grid*, vol. 2, Cardiff, Wales, May 2005, pp. 759–767.
- [171] M. M. Lopez, E. Heymann, and M. A. Senar, "Analysis of dynamic heuristics for workflow scheduling on grid systems," in *The Fifth International Symposium on Parallel and Distributed Computing*, Timisoara, Romania, Jul. 2006, pp. 199–207.
- [172] J. Zhao, H. Li, C. Wu, Z. Li, Z. Zhang, and F. Lau, "Dynamic pricing and profit maximization for the cloud with geo-distributed data centers," in *IEEE INFOCOM*, Toronto, Canada, May 2014, pp. 118–126.
- [173] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, Oct. 2010.
- [174] S. Tang, J. Yuan, and X.-Y. Li, "Towards optimal bidding strategy for amazon ec2 cloud spot instance," in *IEEE International Conference on Cloud Computing (CLOUD)*, Honolulu, Hawaii, USA, Jun. 2012, pp. 91–98.
- [175] D. Poola, K. Ramamohanarao, and R. Buyya, "Fault-tolerant workflow scheduling using spot instances on clouds," *Procedia Computer Science*, vol. 29, no. 1, pp. 523–533, Dec. 2014.
- [176] S. Shen, K. Deng, A. Iosup, and D. Epema, "Scheduling jobs in the cloud using on-demand and reserved instances," in *Euro-Par Parallel Processing*. Aachen, Germany: Springer, Aug. 2013, pp. 242–254.
- [177] S. Yi, D. Kondo, and A. Andrzejak, "Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud," in *IEEE 3rd International Conference on Cloud Computing (CLOUD)*, Miami, Florida, USA, Jul. 2010, pp. 236–243.
- [178] S. Bardhan and D. Milojevic, "A mechanism to measure quality-of-service in a federated cloud environment," in *Proceedings of the workshop on Cloud services, federation, and the 8th open cirrus summit*. San Jose, CA, USA: ACM, Sep. 2012, pp. 19–24.
- [179] W. Qiang, X. Zheng, and C.-H. Hsu, *Cloud Computing and Big Data*. Springer, 2016, vol. 9106.
- [180] J. B. Abdo, J. Demerjian, H. Chaouchi, K. Barbar, and G. Pujolle, "Cloud federation means cash," in *Third International Conference on e-Technologies and Networks for Development (ICeND)*, Beirut, Lebanon, Apr. 2014, pp. 39–42.
- [181] A. N. Toosi, R. N. Calheiros, R. K. Thulasiram, and R. Buyya, "Resource provisioning policies to increase iaas provider's profit in a federated cloud environment," in *IEEE 13th International Conference on High Performance Computing and Communications (HPCC)*, Alberta, Canada, Sep. 2011, pp. 279–287.
- [182] M. Hadji and D. Zeglache, "Mathematical programming approach for revenue maximization in cloud federations," *IEEE Transactions on Cloud Computing*, vol. 1, no. 99, pp. 1–14, Feb. 2015.
- [183] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica: Journal of the Econometric Society*, vol. 28, no. 3, pp. 497–520, Nov. 1960.
- [184] G. Darzanos, I. Koutsopoulos, and G. D. Stamoulis, "A model for evaluating the economics of cloud federation," in *IEEE 4th International Conference on Cloud Networking (CloudNet)*, Niagara Falls, Canada, Oct. 2015, pp. 291–296.
- [185] J. Abate and W. Whitt, "Transient behavior of the m/m/l queue: Starting at the origin," *Queueing Systems*, vol. 2, no. 1, pp. 41–65, Mar. 1987.
- [186] M. Aazam and E.-N. Huh, "Advance resource reservation and qos based refunding in cloud federation," in *IEEE Globecom Workshops*, Austin, TX, USA, Dec. 2014, pp. 139–143.
- [187] M. Aazam and E.-N. Huh, "Broker as a service (baas) pricing and resource estimation model," in *IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom)*, Singapore, Dec. 2014, pp. 463–468.
- [188] M. Aazam and E.-N. Huh, "Fog computing micro datacenter based dynamic resource estimation and pricing model for iot," in *IEEE 29th International Conference on Advanced Information Networking and Applications*, Guwangiu, Korea, Mar. 2015, pp. 687–694.
- [189] J. Altmann and M. M. Kashef, "Cost model based service placement in federated hybrid clouds," *Future Generation Computer Systems*, vol. 41, no. 1, pp. 79–90, Dec. 2014.
- [190] Z. Zheng, Y. Gui, F. Wu, and G. Chen, "Star: strategy-proof double auctions for multi-cloud, multi-tenant bandwidth reservation," *IEEE Transactions on Computers*, vol. 64, no. 7, pp. 2071–2083, Aug. 2015.

- [191] J. Proakis, *Digital Communications*, ser. Electrical engineering series. McGraw-Hill, 2001. [Online]. Available: <https://books.google.com.sg/books?id=sbr8QwAACAAJ>
- [192] T. K. Forde, I. Macaluso, and L. E. Doyle, "Exclusive sharing & virtualization of the cellular network," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Aachen, Germany, May 2011, pp. 337–348.
- [193] T. Sandholm, "Algorithm for optimal winner determination in combinatorial auctions," *Artificial intelligence*, vol. 135, no. 1, pp. 1–54, Feb. 2002.
- [194] S. Misra, S. Das, M. Khatua, and M. S. Obaidat, "Qos-guaranteed bandwidth shifting and redistribution in mobile cloud environment," *IEEE Transactions on Cloud Computing*, vol. 2, no. 2, pp. 181–193, Dec. 2014.
- [195] X. Wang, Z. Li, P. Xu, Y. Xu, X. Gao, and H.-H. Chen, "Spectrum sharing in cognitive radio networks—an auction-based approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 3, pp. 587–596, Dec. 2010.
- [196] S. Das, M. Khatua, S. Misra, and M. Obaidat, "Quality-assured secured load sharing in mobile cloud networking environment," *IEEE Transactions on Cloud Computing*, to appear.
- [197] X. Wang. (2010) C-ran: the road towards green ran. [Online]. Available: <http://labs.chinamobile.com/cran/wp-content/uploads/CRAN/relax%20@%20under%20line%20relax%20relaxwhite%20relax%20@%20under%20line%20relaxpaper%20relax%20@%20under%20line%20relaxv2%20relax%20@%20under%20line%20relax%20relax5%20relax%20@%20under%20line%20relaxEN.pdf>
- [198] V. N. Ha, L. B. Le, and N.-D. Dao, "Energy-efficient coordinated transmission for cloud-rans: Algorithm design and trade-off," in *48th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.
- [199] O. Dhifallah, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Decentralized group sparse beamforming for multi-cloud radio access networks," in *IEEE GLOBECOM*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
- [200] H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Distributed cloud association in downlink multicloud radio access networks," in *49th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, Mar. 2015, pp. 1–3.
- [201] P. C. Chu and J. E. Beasley, "A genetic algorithm for the multidimensional knapsack problem," *Journal of heuristics*, vol. 4, no. 1, pp. 63–86, Jun. 1998.
- [202] H. Kellerer, U. Pferschy, and D. Pisinger, "Knapsack problems. 2004."
- [203] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal *et al.* (2014) Mobile-edge computing introductory technical white paper. [Online]. Available: <https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge/relax%20@%20under%20line%20relaxcomputing%20relax%20@%20under%20line%20relax-introductory%20relax%20@%20under%20line%20relaxtechnical%20relax%20@%20under%20line%20relaxwhite%20relax%20@%20under%20line%20relaxv118-09-14.pdf>
- [204] S. Wang, K. Chan, R. Uргаonkar, T. He, and K. K. Leung, "Emulation-based study of dynamic service placement in mobile micro-clouds," in *IEEE Military Communications Conference*, Tampa, FL, USA, Oct. 2015, pp. 1046–1051.
- [205] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, Oct. 2009.
- [206] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *IEEE Computer*, vol. 43, no. 4, pp. 51–56, Apr. 2010.
- [207] J. Zhang, T. Xiong, and W. Lou, "Community clinic: Economizing mobile cloud service cost via cloudlet group," in *IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Pennsylvania, USA, Oct. 2014, pp. 208–216.
- [208] J. Chen, X. Chen, and X. Song, "Bidder's strategy under group-buying auction on the internet," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 32, no. 6, pp. 680–690, Nov. 2002.
- [209] A. Jin, W. Song, P. Wang, D. Niyato, P. Ju *et al.*, "Auction mechanisms toward efficient resource sharing for cloudlets in mobile cloud computing," *IEEE Transactions on Services Computing*, to appear.
- [210] A. Jin, W. Song, W. Zhuang *et al.*, "Auction-based resource allocation for sharing cloudlets in mobile cloud computing," *IEEE Transactions on Emerging Topics in Computing*, to appear.
- [211] F. Teng and F. Magoules, "Resource pricing and equilibrium allocation policy in cloud computing," in *IEEE 10th International Conference on Computer and Information Technology (CIT)*, Bradford, United Kingdom, Jul. 2010, pp. 195–202.
- [212] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 641–656, Nov. 2014.
- [213] Y. Wu and L. Ying, "A cloudlet-based multi-lateral resource exchange framework for mobile users," in *IEEE INFOCOM*, Bradford, United Kingdom, Jul. 2015, pp. 927–935.
- [214] D. P. Williamson, "The primal-dual method for approximation algorithms," *Mathematical Programming*, vol. 91, no. 3, pp. 447–478, Feb. 2002.
- [215] H. Uzawa, "Walras' tatonnement in the theory of exchange," *The Review of Economic Studies*, vol. 27, no. 3, pp. 182–194, Jun. 1960.
- [216] S. Y. Vaezpour, K. Wu, and G. C. Shoja, "Mobile telecom cloud brokerage with orchestrated multi-tier resource pooling," in *IEEE 4th International Conference on Cloud Networking (CloudNet)*, Niagara Falls, Canada, Oct. 2015, pp. 146–152.
- [217] P. Raghavan and C. D. Tompson, "Randomized rounding: a technique for provably good algorithms and algorithmic proofs," *Combinatorica*, vol. 7, no. 4, pp. 365–374, Dec. 1987.
- [218] D. P. Anderson, "Boinc: A system for public-resource computing and storage," in *Fifth IEEE/ACM International Workshop on Grid Computing*, Pennsylvania, USA, Nov. 2004, pp. 4–10.
- [219] S. Di, C.-L. Wang, L. Cheng, and L. Chen, "Social-optimized win-win resource allocation for self-organizing cloud," in *International Conference on Cloud and Service Computing (CSC)*, Hong Kong, China, Dec. 2011, pp. 251–258.
- [220] J.-P. Sheu, S.-C. Tu, and C.-H. Yu, "A distributed query protocol in wireless sensor networks," *Wireless Personal Communications*, vol. 41, no. 4, pp. 449–464, Jun. 2007.
- [221] T. Mukherjee, P. Dutta, V. G. Hegde, and S. Gujar, "Risc: Robust infrastructure over shared computing resources through dynamic pricing and incentivization," in *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Hyderabad, India, May 2015, pp. 1107–1116.
- [222] K. S. Dilip, N. Sadashiv, and R. Goudar, "Priority based resource allocation and demand based pricing model in peer-to-peer clouds," in *International Conference on Advances in Computing, Communications and Informatics*, Delhi, India, Sep. 2014, pp. 1210–1216.
- [223] K. Danniswara, H. P. Sajjad, A. Al-Shishtawy, and V. Vlassov, "Stream processing in community network clouds," in *International Conference on Future Internet of Things and Cloud (FiCloud)*, Rome, Italy, Aug. 2015, pp. 800–805.
- [224] A. M. Khan, X. Vilaça, L. Rodrigues, and F. Freitag, "Towards incentive-compatible pricing for bandwidth reservation in community network clouds," in *International Conference on Economics of Grids, Clouds, Systems, and Services (GECON'15)*, Cluj-Napoca, Romania, Sep. 2015, pp. 251–264.
- [225] X. Chu, H. Liu, Y.-W. Leung, Z. Li, and M. Lei. (2013) User-assisted cloud storage system: Opportunities and challenges. [Online]. Available: <http://pages.cpsc.ucalgary.ca/~zongpeng/publications/mmte/relax%20@%20under%20line%20relax2013.pdf>
- [226] J. Zhao, X. Chu, H. Liu, Y.-W. Leung, and Z. Li, "Online procurement auctions for resource pooling in client-assisted cloud storage systems," in *IEEE Conference on Computer Communications (INFOCOM)*, Hong Kong, China, May 2015, pp. 576–584.
- [227] A. Davoli and A. Mei, "Triton: a peer-assisted cloud storage system," in *Proceedings of the First Workshop on Principles and Practice of Eventual Consistency*. Amsterdam, Netherlands: ACM, Apr. 2014, pp. 4–10.
- [228] S. Di and C.-L. Wang, "Dynamic optimization of multiattribute resource allocation in self-organizing clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 3, pp. 464–478, May 2013.
- [229] X. Wu, M. Liu, W. Dou, L. Gao, and S. Yu, "A scalable and automatic mechanism for resource allocation in self-organizing cloud," *Peer-to-Peer Networking and Applications*, vol. 9, no. 1, pp. 28–41, Jan. 2016.
- [230] P. Khethavath, J. Thomas, E. Chan-Tin, and H. Liu, "Introducing a distributed cloud architecture with efficient resource discovery and optimal resource allocation," in *Ninth World Congress on Services (SERVICES)*, Santa Clara Marriott, CA, USA, Jul. 2013, pp. 386–392.

- with a novel pricing scheme," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 2, pp. 556–569, Apr. 2009.
- [275] G. Nan, Z. Mao, M. Li, Y. Zhang, S. Gjessing, H. Wang, and M. Guizani, "Distributed resource allocation in cloud-based wireless multimedia social networks," *IEEE Network*, vol. 28, no. 4, pp. 74–80, Jul. 2014.
- [276] G. Nan, C. Zang, R. Dou, and M. Li, "Pricing and resource allocation for multimedia social network in cloud environments," *Knowledge-Based Systems*, vol. 88, no. 1, pp. 1–11, Nov. 2015.
- [277] A. Ekert, "Complex and unpredictable cardano," *International Journal of Theoretical Physics*, vol. 47, no. 8, pp. 2101–2119, Aug. 2008.
- [278] L. E. Li, Z. M. Mao, and J. Rexford, "Toward software-defined cellular networks," in *European Workshop on Software Defined Networking (EWSNDN)*, Darmstadt, Germany, Oct. 2012, pp. 7–12.
- [279] C. Chaudet and Y. Haddad, "Wireless software defined networks: Challenges and opportunities," in *IEEE International Conference on Microwaves, Communications, Antennas and Electronics Systems (COMCAS)*, Tel Aviv, Israel, Oct. 2013, pp. 1–5.
- [280] N. A. Jagadeesan and B. Krishnamachari, "Software-defined networking paradigms in wireless networks: a survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 27, Jan. 2015.
- [281] C. Bernardos, A. La Oliva, P. Serrano, A. Banchs, L. M. Contreras, H. Jin, and J. C. Zúñiga, "An architecture for software defined wireless networking," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 52–61, Jun. 2014.
- [282] W. Dunham. (1990) Cardano and the solution of the cubic. [Online]. Available: <http://www.ms.uky.edu/~corso/teaching/math330/Cardano.pdf>
- [283] (2015) Ip flow mobility and seamless wireless local area network (wlan) offload. 3GPP TS 23.261. [Online]. Available: <http://www.3gpp.org/DynaReport/23261.htm>
- [284] S. Kang and W. Yoon, "Sdn-based resource allocation for heterogeneous lte and wlan multi-radio networks," *The Journal of Supercomputing*, vol. 72, no. 4, pp. 1342–1362, Apr. 2016.
- [285] Y. Zhang, L. Liu, Y. Gu, D. Niyato, M. Pan, and Z. Han, "Offloading in software defined network at edge with information asymmetry: A contract theoretical approach," *Journal of Signal Processing Systems*, vol. 83, no. 2, pp. 1–13, May 2016.
- [286] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 2001–2014, Sep. 2013.
- [287] H. Eghbali and V. W. Wong, "Bandwidth allocation and pricing for sdn-enabled home networks," in *IEEE ICC*, London, United Kingdom, Jun. 2015, pp. 5342–5347.
- [288] L. Zhang, W. Wu, and D. Wang, "Time dependent pricing in wireless data networks: Flat-rate vs. usage-based schemes," in *IEEE INFOCOM*, Toronto, ON, Canada, May 2014, pp. 700–708.
- [289] T. Feng, J. Bi, and K. Wang, "Joint allocation and scheduling of network resource for multiple control applications in sdn," in *Network Operations and Management Symposium (NOMS)*, Krakow, Poland, May 2014, pp. 1–7.
- [290] F. Tao, B. Jun, and W. Ke, "Allocation and scheduling of network resource for multiple control applications in sdn," *Communications, China*, vol. 12, no. 6, pp. 85–95, Jun. 2015.
- [291] M. Yokoo, Y. Sakurai, and S. Matsubara, "The effect of false-name bids in combinatorial auctions: New fraud in internet auctions," *Games and Economic Behavior*, vol. 46, no. 1, pp. 174–188, Jan. 2004.
- [292] M. Yokoo, Y. Sakurai, and S. Matsubara, "Robust double auction protocol against false-name bids," *Decision Support Systems*, vol. 39, no. 2, pp. 241–252, Apr. 2005.
- [293] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman, "Creating an ambient-intelligence environment using embedded agents," *IEEE Intelligent Systems*, vol. 19, no. 6, pp. 12–20, Nov. 2004.
- [294] H.-T. Zhang, F.-Y. Xu, and L. Zhou, "Artificial neural network for load forecasting in smart grid," in *International Conference on Machine Learning and Cybernetics*, vol. 6, Qingdao, China, Jul. 2010, pp. 3200–3205.
- [295] B. Li, S. Gangadhar, S. Cheng, and P. K. Verma, "Predicting user comfort level using machine learning for smart grid environments," in *IEEE Innovative Smart Grid Technologies (ISGT)*, Anaheim, CA, USA, Jan. 2011, pp. 1–6.
- [296] R. Bogacz, "Optimal decision-making theories: linking neurobiology with behaviour," *Trends in cognitive sciences*, vol. 11, no. 3, pp. 118–125, Mar. 2007.
- [297] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, Nov. 2010, pp. 267–280.
- [298] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 63–74, Oct. 2008.
- [299] G. He. (2002) Destination-sequenced distance vector (dsv) protocol. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc>
- [300] T. Clausen and P. Jacquet. (2003) Optimized link state routing protocol (olsr). [Online]. Available: <https://www.rfc-editor.org/rfc/rfc3626.txt>
- [301] F. Li, J. Wan, P. Zhang, D. Li, D. Zhang, and K. Zhou, "Usage-specific semantic integration for cyber-physical robot systems," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 15, no. 3, pp. 50–59, Jul. 2016.
- [302] G. Hu, W. P. Tay, and Y. Wen, "Cloud robotics: architecture, challenges and applications," *IEEE Network*, vol. 26, no. 3, pp. 21–28, May 2012.
- [303] L. Wang, M. Liu, and M. Q.-H. Meng, "Hierarchical auction-based mechanism for real-time resource retrieval in cloud mobile robotic system," in *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, Jun. 2014, pp. 2164–2169.
- [304] L. Wang and M. Q.-H. Meng, "A game theoretical bandwidth allocation mechanism for cloud robotics," in *World Congress on Intelligent Control and Automation (WCICA)*, Beijing, China, Jul. 2012, pp. 3828–3833.
- [305] L. Wang, M. Liu, and M. Q.-H. Meng, "A pricing mechanism for task oriented resource allocation in cloud robotics," in *Robots and Sensor Clouds*. Springer, 2016, pp. 3–31.
- [306] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De 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, May 2015.
- [307] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muramidhar, P. Polakos, V. Srinivasan, and T. Woo, "Cloudiq: a framework for processing base stations in a data center," in *Proceedings of the 18th annual international conference on Mobile computing and networking*. Istanbul, Turkey: ACM, Aug. 2012, pp. 125–136.
- [308] J. Elias, F. Martignon, S. Paris, and J. Wang, "Optimization models for congestion mitigation in virtual networks," in *IEEE 22nd International Conference on Network Protocols (ICNP)*, Triangle Park, North Carolina, USA, Oct. 2014, pp. 471–476.
- [309] J. Elias, F. Martignon, S. Paris, and J. Wang, "Efficient orchestration mechanisms for congestion mitigation in nfv: Models and algorithms," *IEEE Transactions on Services Computing*, to appear.