A Survey of Pricing Aware Traffic Engineering in Cloud Computing

Xiaocui Sun[1,4], Xinran Zhuo[2], Zhijun Wang[3]

[1] Department of Computer Science, GuangDong Pharmaceutical University, China
[2] Glasgow College, University of Electronic Science and Technology of China, China
[3] Department of Computer Science, The University of Texas at Arlington, USA
[4] Medicinal Information & Real World Engineering Technology Center
xiaocuisun1002@hotmail.com, 2017200604022@std.uestc.edu.cn, zhijun.wang@uta.edu

Abstract

Cloud computing is the on-demand availability of computer resources and has drawn widely attention from industry. How to charge cloud resource usage to satisfy the customers' requirements while considering a variety of parties' interest is becoming increasingly important. Fair and effective pricing aware traffic engineering model could improve the resource utilization, attract more users and increase revenues for service providers. In this paper, we presents the basic core components and give a comparative review of latest and most appropriate pricing aware traffic engineering including network utility maximization based traffic control, game theory approaching bandwidth allocation and so on in cloud computing. The purpose of this paper is to bring out the important characteristics of pricing models with traffic engineering instead of survey all existing solutions. We hope to provide readers with a wide range of options and factors for designing better models in the future.

Keywords: Cloud computing, Traffic engineering, Pricing models, Economic models

1 Introduction

Bandwidth is valuable and precious resources in today's data center networks, such as Alibaba, Amazon and Google clouds. These service providers aim to provide attractive cloud services (e.g. data storage, data access, data computing etc.) to their users. Most of current service providers charge their users usually in a fixed pricing unit according to their consumption of CPU and memory resources, but, independent on the network bandwidth usage. One critical reason is due to the fluctuated and highly dynamic nature of the network condition and makes it hard to provide bandwidth guarantee. The competition of intensive bandwidth resource especially during the peak time may result in unpredictable performance on task response times, which will harm the users' satisfaction and may lose high payment users, which eventually results in the loss of cloud provider's profit. Actually, many customers would like to pay more money to obtain their desired performance, for example, a video player expects a smooth video play and can pay more to get such service [1]. Hence, it is emergency technology to develop pricing models based on the integrating both computing resource and networking resource utilization.

A good price model integrating both computation and network bandwidth usage should have the following features. Firstly, the network bandwidth allocation should be proportional to the prices that the users paid during the network bandwidth shortage. Secondly, the price model should be easily implemented in the existing data center and has good scalability. Thirdly, the price model should not degenerate the overall network performance. Finally, the pricing modeling should allow different users to select their desired price model to maximize their requirements while minimize their costs.

Some of the previous research have reviewed the price modeling in cloud computing. Samimi and Patel [2] reviewed and compared the economic models in grid and cloud computing. Luong et al. [3] presented a thorough comparison between many proposed cloud computing pricing models and schemes by considering many factors that affect pricing and user satisfaction. However, they did not focus on the traffic engineering based pricing model. This paper brings out the basic core principles of fair and effective pricing aware traffic engineering. Besides, a comparative review of the latest and most appropriate pricing models with traffic engineering is presented. The comparison is based on many aspects such as provider's profit vs. social welfare, static pricing vs. dynamic pricing, central control vs. distributed management, intra traffic vs. inter traffic. Challenges for integrating pricing model and traffic control are also discussed to help

[*] Corresponding Author: Zhijun Wang; E-mail: zhijun.wang@uta.edu

DOI: 10.3966/160792642020032102004

users to propose better models for the future.

The remainder of the paper is organized as follows. Section 2 present the overview of cloud computing. Section 3 gives a brief description of traffic properties and the objective of price related traffic engineering. Section 4 describes different data center pricing aware traffic engineering models. Section 5 discusses challenges for integrating pricing model and traffic control. And we conclude our work in Section 6.

2 Cloud Datacenter Network

Cloud computing provides services for the consumers to share software, computing and network resources via the Internet in distributed environments. This resourcing sharing service model can greatly increase the resource utilization and then reduce customer costs to run their services. The cloud services contain different elements, features, functions and dimensions covering many areas, including:

- IaaS: Infrastructure as a Service is a model in which customers lease equipment such as physical or virtual machines (VMs) through the internet instead of purchasing hardware from service providers. The price typically charged on a per-use basis based on the amount of computing resources allocated to the customers.
- PaaS: Platform as a Service is a model whereby customers can lease a computing platform includes operating systems, hardware, programming language, servers, and databases to develop and run applications through the Internet.
- SaaS: Software as a Service model is where service providers install and maintain software applications in the cloud and the customers pay for software usage and upgrades.

Cloud computing contains six layers including Cloud System Resources Fabric, Communication Connectivity, Unified Resources, Collective and Composite, Middleware and Application as shown in Figure 1.

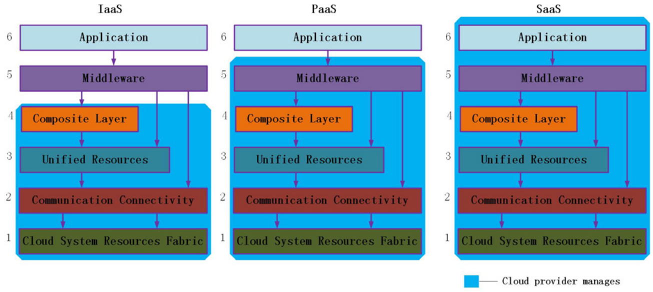

Figure 1. Cloud computing architecture and three types of services

Datacenter networks are used as the majority infrastructure to host cloud services including special computation and storage resources. A data center is composed by a crowd of networked computer servers which are responsible for remote storage, huge data processing and distribution [4]. The aim of data center network is to build a robust network that could provide low latency (e.g. hundreds of microseconds) and high bandwidth across servers. Recently, alternative cloud infrastructure models using multiple data centers from many providers have been discussed [5]. This kind of network could overcome several shortcomings of the single data center structure, such as single point failures, huge energy consuming, and geographically distributed data. Cloud data center networking can be divided into intra and inter data center networking.

Figure 2 illustrates a data center network with a large group of racks which are fundamentally groups of machines connected to a Top of Rack (ToR) switch. Up to hundreds of racks could form a cluster [4]. Large Cluster interconnection connected ToR switches so that across racks communication is provided. Many such clusters with thousands of machines per cluster compose a data center. Ideally, the network should act as a large non-blocking switch, with all servers connected directly to the switch, allowing them to communicate simultaneously at maximum speed.

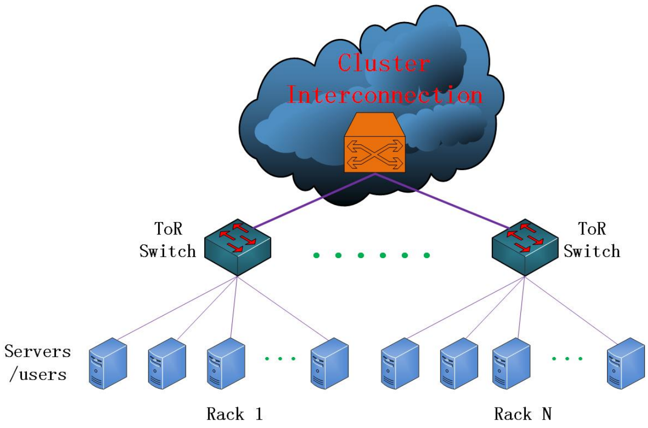

Figure 2. A datacenter structure

3 Objectives of Price Aware Traffic Engineering

We discuss the objectives of price related traffic engineering in this section.

Network utilization maximization. The network utility maximization (NUM) problem can be stated as follows,

$$\max \sum_{i}^{N} U_i(x_i) \qquad \textbf{(1)}$$

$U_i(x_i)$ is any given concave utility function, for user flow i with flow rate x_i for $i = 1, 2, \ldots, N$, where N is the total number of active flows. Both the link bandwidth constraints for all links in the network and the minimum flow rate constraints for inelastic flows, i.e., flows with minimum rate, low latency flows or flows with deadline requirements. Different users have different utility functions, the user utility function can be reflected into the user pricing model. A family of fully distributed optimal traffic control laws can be derived using any given concave user utility function. A given utility function determines the traffic control laws. Hence, different utility function results in different traffic control. A customer with the traffic control laws combined with pricing could maximize the total network utility.

Price based fairness for network bandwidth utilization. Datacenter resources were shared by many applications which generated many flows. Fair share of network resources according to their Service Level Agreement (SLA) mitigates the congestion problem and prevents offensive behavior. CPU cycles, memory performance and network bandwidth are all significant metrics which severely impact the job completion. The first two metrics have been standardized as the quantifiable VMs performance metrics. However, data center network bandwidth, which can seriously affect job completion time, has not considered in the price modeling. Fairness criteria should determine how link bandwidth is allocated among individual user. With fair traffic control, users with higher price should occupy more bandwidth than a user with lower price. Users can choose their favored payment based on their job requirement. Peak hour and off-peak hour usage could be balanced.

The fairness could be defined as flow level fairness or VM level fairness. In flow level fairness, the bandwidth is allocated to individual flows. While in VM level fairness, the bandwidth is determined by the number of VMs instead of the number of flows to prevent aggressive users' behavior from obtaining more resources by creating many flows.

Traffic control with price could improve user's satisfaction in mainly two ways. (1) *Flow Completion Time (FCT):* FCT is calculated as the time between the generation and completion of a flow. It determines the response time of an application or request. The end users asking for real-time application could have a good experience and satisfied with the network service if the FCT is low [6]. Queuing and packet loss are two key factors that affect the performance of FCT in data centers. (2) *Deadline guarantee:* some users would like to pay more to guarantee their job response time before a fixed deadline [7]. It is important for the cloud provider to fulfill the SLA of their customers. Traffic control could give priorities to different flows to allocate the timely delivery of flows. The quality of services could be increased as the fraction of jobs finished on time increases.

Profit and social welfare. The profit related resource allocation problem can be expressed as

$$\max_{x} \sum_{i}^{N} U_i(x_i) - C(x) \qquad \textbf{(2)}$$

s.t. $x_i \in [a_i,b_i]$, $i = 1, ..., N$,

Here, C(x) is the cost function which is strictly convex, and a_i and b_i are flow rate the constants. Assume user *i* pays a price p_i for using resource x_i, and p = (p_1... p_N), is a price vector. Now it can be expressed as follows:

$$\max_{x \in \pi_i [a_i, b_i]} \sum_{i}^{N} (U_i(x_i) - p_i x_i) + (p^T x - C(x)), \qquad \textbf{(3)}$$

Cloud user *i* wants to improve ($U_i(x_i)$-p_ix_i) to have better surplus for using resource units x_i, and cloud provider intends to increase (p^Tx-$C(x)$) to have a greater profit. Cloud providers' aim is to maximize their profit, while users want to achieve high social welfare which is defined as the sum of user utilities minus a coupled provider cost via resource allocation. It seems like a tradeoff, as buyers want to pay less but sellers want to earn more. However, a reasonable price with satisfied or differential services provided based on the needs of users with different objectives could attract more buyers. By effectively setting an attractive price for both parties and adjusts the traffic control law accordingly, better profits and social welfare could be achieved.

4 Datacenter Pricing Models

Many useful theoretical studies for data center pricing aware traffic engineering have been derived. Game theory approach is very frequently applied in pricing for traffic control. Many different efficient methods are also helpful to give a reasonable price and traffic matching. Bandwidth reservation, bandwidth allocation and traffic scheduling are the methods used very often. In the following, different data center pricing aware traffic engineering are described.

4.1 Provider's Profit vs. Social Welfare

As mentioned before, from the view of providers, good pricing aware traffic engineering should satisfy users' bandwidth requirements and achieve efficient resource utilization to maximize providers' profit, while social welfare maximization means to maximize the surplus of everyone including both providers and users. On the other hand, from the aspect of users, they would like to pay less to enjoy more network bandwidth.

SoftBW [8] implements a usage-based charging pricing model and a fulfillment-based scheduling algorithm to satisfy the bandwidth requirements of users, but also achieve efficient resource utilization to benefit providers' profit. It charges users based on their real bandwidth usage by monitoring the ratio of the used bandwidth to the committed bandwidth guarantee. The central scheduling is used to guarantee bandwidth usage based on different application requirements (e.g. strict rate guarantee and deadline guarantee). It is shown that this new mechanism could achieve fairness when the bandwidth is over-committed while guaranteeing minimum bandwidth under normal bandwidth condition.

The price aware traffic engineering proposed in [9] could achieve both fast distributes resource allocation and utility maximization by only feeds back gradient information. And eventually the users' social welfare could be maximized. The basic idea is to maximize the social welfare. The provider calculates a price vector as a weighted average of the marginal cost under current resource requirements and the previous price based on all the resource allocation requirements from users. Then, each user computes its resource requirement under the new price aims to maximize its surplus, that is, its utility minus the price. However, the gradient information used may give more complexity to the computation of the bandwidth price.

An explicit pricing model [10] is developed to enable cloud bandwidth reservation based on historical workload data even with the presence of demand uncertainty such as burstiness and covariance. The users can pay extra reservation fee to enjoy guaranteed portion of performance and the normal usage fee for the rest demand served with best effort. The new model is tested based on the real-world video workload dataset and operates effectively. This pricing mechanism provide cloud bandwidth reservations with the objective to maximize social welfare whereas the profit of the providers.

Either the provider's or the user interest is well considered in the aforementioned schemes. [11] tries to maximize the time-average profit by considering the revenue from serving the requests and the cost happened because of the usages of bandwidth and computation resources. The bandwidth allocation is determined by the calculated weights of different tasks which also are the basis to provide differentiated services. Firstly, they provide the bandwidth allocation algorithm benefit from the concept of the store-and-forward scheme and hop-by-hop bandwidth allocation. Then they improve this idea by letting data centers to collaboratively decide bandwidth allocation and calculate a forwarding path segment instead of only one hop further for each task. This could alleviate the data-transfer latency and routing loops caused by the first algorithm.

4.2 Static Pricing vs. Dynamic Pricing

The price of using the bandwidth can be statically or dynamically computed. Currently, most data center pricing methods are static pricing. In a static pricing scheme, the resource is fairly allocated based on the users' advanced indicated requests, e.g. the priority class of transfers, exact latency requirements, precise demand deadlines and so on. In order to reflect the real bandwidth usage and satisfy different traffic needs of cloud users, dynamic pricing is proposed adapting to required changes and usage models.

[12] provides minimum bandwidth guarantees for intra-user traffic and certain dependent user and upper-bound bandwidth proportional to the payment. This could guarantee robust network sharing while restricting the chatty users' behavior. Here, one important premise to fairly allocate the resource is that the users should be honest and generous when they indicate their requests.

The solution proposed in [13] targets to achieve three goals: support minimum network bandwidth guarantee, achieve high network utilization and provide network proportionality to price paid. The bandwidth allocated along congested links is based on the ratio of the number of VMs occupied by a user versus the total number of VMs along that link or in the view of the whole subnetwork. Proportional sharing can be achieved by assigning different weights to the source and destination VMs ratio in a tree based topology. However, the price is set only based on statically allocated bandwidth.

The dynamic pricing scheme proposed in [14] is based on a Shapley value based auction. Every user has its own Shapley value which is calculated by the ISP provider based on the users' share of the overall ISP. Users should specify an amount of data and the transmission deadline when they bid their bandwidth. A user's bid price is accepted if it is higher than the Shapley value. In the proposed offline auction version, after an ISP received all the users' bid information, it uses the calculated Shapley value to decide if this user's bid will be accepted or not and then figures out the corresponding optimal traffic schedule afterwards. While in the online auction version, the similar ideas are kept, but the user's request will be response directly instead of waiting all the others. The Shapley value here is calculated based on the estimation simulation or formula. This pricing model pays less attention to profit maximization and resource utilization optimization.

A new framework proposed in [15] not only considers traffic engineering but also online dynamic pricing. A price quote according to different bandwidth guarantee is generated based on the users' demand. Duel pricing, recent advances in combinatorial market design and statistical learning are used to update the price dynamically. The evaluation based on the topology and traffic observed on a large production WAN shown that it can achieve up to 80% of the social welfare compared with previous usage-based alternatives. The solution mainly aims to minimize inter-data center traffic transit costs by alleviating the peak bandwidth usage.

4.3 Central Control vs. Distributed Management

In centralized schemes, a central unit coordinates transmissions in the network to fit the uses' pricing traffic engineering requirement. The global view of network topology and resources, switches' state information, and uses' traffic demands are provided to the central unit. The provider can decide the uses' prices based on the resource allocation requirements from users (e.g. flow sizes, deadlines, priorities and so on) and coordinate the corresponding transmission pattern in a way that optimizes the performance and minimizes contentions. A global view of network resources and flow properties can provide higher performance. But consistent network updates may form a network hot spot in large networks.

Shifting more controls to end users or switches may achieve higher scalability and reliability as easier accessible to information about flows from multiple end users. In distributed scheme, each user could compute its resource requirement under the providers' price by maximizing its surplus. Due to lack of coordination and restricted knowledge of network condition and properties, it is hard to achieve globally optimal.

Uncertainty of three parameters was considered in [16]: the number of VMs required per class, the bandwidth required per VM and the costs of both VMs and bandwidth. To avoid of over-and under-provisioning of VMs and bandwidth resources, they employ stochastic programming to optimally reserve VMs and bandwidth across multiple time stages despite uncertain demand. As the uncertainty of pricing and demand is considered, they also conducted sensitivity analysis to exam the sensitivity to parameter changes, such as changes in cost may result in rerouted traffic. Both providers and users can make decisions about the price on the basis of changes. The computational overhead and the random network delay should be considered when implementing the above methods especially in distributed system.

4.4 Intra Traffic vs. Inter Traffic

Usually, the cloud services are provided by the cloud provider which the users registered in. So, the traffic engineering could be considered inside a cloud provider. However, due to the explosion in scale of modern cloud platforms, it is a common trend for applications running in cloud data centers to use applications and services from other cloud providers. Some of these services are run by users themselves, while others services are offered by the cloud provider that users can employ as application building blocks. The interconnection of applications and services between different clouds leads to an increasing amount of user-user and user-provider communication. It is increasingly important to develop fast and robust pricing aware traffic engineering solutions to meet the requirements of inter-tenant traffic applications.

Many of the approaches mentioned previously are concentrated on intra Traffic. Inter data center network bandwidth pricing is also widely studied recently [14-15, 17-18]. In [17], it mainly achieves three goals for allocating bandwidth to inter-data center traffic: bandwidth guarantee, minimizing the total network cost, and avoiding potential bottleneck problem at low cost links. A distributed algorithm is proposed based on the auxiliary variable method and the alternating direction method of multipliers (ADMM) [11] to select the cost-effective paths. The path selection does not consider the time factor which may lead to comparatively long communication delays during transmission. So, it is not suitable for time-sensitive interactive traffic.

[18] also focused on inter-data center network where traditional application providers can reserve bandwidth from Internet providers like Google and Microsoft to guarantee their WAN traffic. A two-stage Stackelberg game is used to model the interaction between Internet providers and application providers. A cooperated Nash bargaining game among Internet providers is used to decide the price and the corresponding allocated bandwidth. Then, a non-cooperated game among application providers will perform to decide the individual bandwidth reserved from different Internet providers. The bandwidth price is calculated based on the geometrical Nash bargaining and demand segmentation method. To benefit both the Internet providers and application providers, the per unit bandwidth price is set in an area that is between the highest and the lowest. And the application providers will benefit if they buy more bandwidth from an Internet provider. One disadvantage of auction based price is limited flexibility. As the consulted price is based on the description of their requested resource demands, it is not easy for the users to change their specified demands, such as utility, duration, path options etc., after the deal is settled. Another disadvantage is potentially longer waiting time that is used for the cloud providers to carefully consider the price for all the users. Table 1 give a summary of recent pricing model with traffic engineering in cloud data center networking.

5 Challenges for Integrating Pricing Model and Traffic Control

Although the existing schemes are feasible, they also have some shortcomings. Firstly, many of these schemes force users to wait, possibly until the deadline, to determine their price and their traffic routing path. Secondly, these schemes often require users to provide detailed descriptions of their requests, including all possible routing, duration, network utility, etc. Therefore, it may lead to slow convergence, unstable equilibrium, or unreliable descriptions. Centralized network traffic control technology can improve network utilization and clear guarantee deadlines without affecting low latency sensitive traffic. These

Table 1. A summary of recent pricing model with traffic enineering in cloud datacenter networking

Pricing Model	Traffic Engineering	Objective	Disadvantage
Usage-based charging [8]	Central scheduling	• Fairness • Over-committed • Minimum bandwidth guaranteeing	More failure
Gradient method [9]	Network resource allocation	• Fast distribute resource allocation • Utility maximization • Social welfare maximization	Computation complexity
Explicit pricing [10]	Bandwidth reservation	Social welfare maximization	No consider the profit of the providers
Smart pricing(weights based) [11]	Bandwidth allocation	Time-average profit maximization	Potentially data-transfer latency
New network sharing framework [12]	Bandwidth guarantee	• Minimum bandwidth guarantees • Upper-bound bandwidth proportional to its payment	Users' behavior dependence
Smart pricing (weights based) [13]	Bandwidth allocation	• Minimum network bandwidth guarantee • Network utilization advance • Network proportionality	Statically allocated bandwidth
Shapley value based auction [14]	Traffic schedule	Optimal traffic schedule	• Less profit maximization • Less resource utilization optimization.
Online dynamic pricing [15]	Bandwidth guarantee	• Social welfare maximization • Inter-datacenter traffic transit costs minimization	Peak bandwidth usage alleviating mainly
Stochastic programming [16]	Bandwidth reservation	• Uncertain demand suitable • Bandwidth reserve optimization	• Computational overhead • Random network delay
Auxiliary variable method and the alternating direction method [17]	Bandwidth guarantee	• Bandwidth guarantee • Minimum total network cost • Potential bottleneck problem avoiding	Comparatively long communication delays
Nash bargaining game [18]	Bandwidth reservation	Modest bandwidth price	• limited flexibility • Potentially longer waiting time

technologies depend critically on detailed traffic information: request priority, precise delay requirements, and transmission deadlines. These functions also need to be provided by users, but simply requesting this information can produce unexpected results. In order to get better service, users may increase priority or shorten deadlines for their own benefit, which will significantly reduce the overall performance of the system. The future pricing model should be embodied in the following four aspects.

Firstly, integrating flow control and pricing. A good traffic control protocol should be able to alleviate the phenomenon of overload through price adjustment and plan the rational utilization of network bandwidth. Through flexible pricing scheme, network bandwidth can be protected from the impact of strategic users, and bandwidth utilization can be improved. Providing a lower price for flexible requests (e.g., requests with a flexible deadline will be charged a lower price than similar requests with a strict deadline). Traffic control management is carried out for different prices to achieve the rational utilization of network bandwidth. Traffic control is an important tool for setting the right price; if paths and traffic times are not carefully planned, some links may take longer or less time than they should [19].

Secondly, providing service deadline guarantee, i.e., minimum rate guarantee. Public cloud providers could charge users based on their network traffic. Generally speaking, traffic leaving the WAN (to the Internet) charges more than traffic within the WAN. More discounts will give to users with flexible deadlines. So that cloud providers could use the network bandwidth more efficiently to provide network bandwidth guarantees or data transmission guarantees within deadlines.

Thirdly, achieving overall network maximization. Despite the existence of network traffic algorithms that meet user deadlines, there is no viable way to motivate customers to report real needs or deadlines. Therefore, even if some users have flexible deadlines, the lack of pricing models cannot motivate users to shift their load to non-busy periods, which can lead to excessive network capacity consumption. Hence, in current network, more capacity should be provided to meet peak demand which will degrade the profit of the provider and affect the satisfaction of the users.

Fourthly, meeting differentiated service requirements. Various new technologies accelerate the divergence of service demands [20]. Network traffic bandwidth can be a key difference between cloud computing data services among different providers. According to different users' different requirements for tasks, different levels of services should be provided. Users with high transmission speed and large demand for traffic bandwidth can pay relatively high prices. Realize different requirements for network bandwidth. Users who pay a higher price should be able to enjoy better bandwidth services. Service level differentiation should be realized in the future network pricing model.

6 Conclusion

This paper reviewed the cloud data center pricing aware traffic engineering. Firstly, we presented cloud computing key concepts and the common architecture of cloud data center. Then we described the special properties of cloud data center traffic and provided the objectives of using traffic engineering in pricing model. After that, many related pricing models were discussed in detail to show the different design concentrations and applied traffic engineering methods. Finally, Challenges for integrating price model and traffic control are discussed. It is obvious that no single particular model could satisfy all the possible criteria because of the conflict of parties' interest and various objectives of different enterprises. Carefully considering a variety of aspects, better pricing aware traffic engineering could achieve higher revenue, more efficient resource utilization and better network performance in the future.

Acknowledgments

This work is supported by China Scholarship Council project (No. 201708440547) and GDPU Higher education innovation strong school project (No. 51359055).

References

[1] X. Liu, Y. Liu, P. Wang, C.-F. Lai, H.-C. Chao, An Adaptive Mode Decision Algorithm Based on Video Texture Characteristics for HEVC Intra Prediction, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 27, No. 8, pp. 1737-1748, August, 2017.

[2] P. Samimi, A. Patel, Review of Pricing Models for Grid and Cloud Computing, *IEEE Symposium on Computers and Informatics*, Kuala Lumpur, Malaysia, 2011, pp. 634-639.

[3] N. C. Luong, P. Wang, D. Niyato, Y. Wen, Z. Han, Resource Management in Cloud Networking Using Economic Analysis and Pricing Models: A Survey, *IEEE Communications Surveys and Tutorials*, Vol. 19, No. 2, pp. 954-1001, Second Quarter, 2017.

[4] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, H. Liu, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, A. Vahdat, Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network, *Communications of the ACM*, Vol. 59, No. 9, pp. 88-97, September, 2016.

[5] Y. Mansouri, A. N. Toosi, R. Buyya, Data Storage Management in Cloud Environments: Taxonomy, Survey, and Future Directions, *ACM Computing Surveys (CSUR)*, Vol. 50, No. 6, pp. 91:1-91:51, January, 2018.

[6] X. Liu, Y. Li, D. Liu, P. Wang, L. T. Yang, An Adaptive CU Size Decision Algorithm for HEVC Intra Prediction Based on Complexity Classification Using Machine Learning, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 1, pp. 144-155, January, 2019.

[7] C. Dai, X. Liu, J. Lai, P. Li, H.-C. Chao, Human Behavior Deep Recognition Architecture for Smart City Applications in the 5G Environment, *IEEE Network*, Vol. 33, No. 5, pp. 206-211, September-October, 2019.

[8] J. Guo, F. Liu, T. Wang, J.-C. Lui, Pricing Intra-datacenter Networks with Over-committed Bandwidth Guarantee, *2017 USENIX Annual Technical Conference (USENIX ATC' 17)*, Santa Clara, CA, USA, 2017, pp. 69-81.

[9] D. Niu, B. Li, An Efficient Distributed Algorithm for Resource Allocation in Large-scale Coupled Systems, *IEEE INFOCOM*, Turin, Italy, 2013, pp. 1501-1509.

[10] D. Niu, C. Feng, B. Li, Pricing Cloud Bandwidth Reservations under Demand Uncertainty, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40, No. 1, pp. 151-162, June, 2012.

[11] W. Lu, P. Lu, Q. Sun, S. Yu, Z. Zhu, Profit-Aware Distributed Online Scheduling for Data-Oriented Tasks in Cloud Datacenters, *IEEE Access*, Vol. 6, pp. 15629-15642, February, 2018.

[12] W. Li, K. Li, D. Guo, G. Min, H. Qi, J. Zhang, Cost-Minimizing Bandwidth Guarantee for Inter-Datacenter Traffic, *IEEE Transactions on Cloud Computing*, Vol. 7, No. 2, pp. 483-494, April-June, 2019.

[13] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, I. Stoica, FairCloud: Sharing the Network in Cloud Computing, *ACM SIGCOMM Computer Communication Review*, Vol. 42, No. 4, pp. 187-198, October, 2012.

[14] W. Shi, C. Wu, Z. Li, A Shapley-value Mechanism for Bandwidth on Demand between Datacenters, *IEEE Transactions on Cloud Computing*, Vol. 6, No. 1, pp. 19-32, January-March, 2018.

[15] V. Jalaparti, I. Bliznets, S. Kandula, B. Lucier, I. Menache, Dynamic Pricing and Traffic Engineering for Timely Inter-datacenter Transfers, *2016 ACM SIGCOMM Conference*, Florianopolis, Brazil, 2016, pp.73-86.

[16] J. Chase, D. Niyato, Joint Optimization of Resource Provisioning in Cloud Computing, *IEEE Transactions on Services Computing*, Vol. 10, No. 3, pp. 396-409, May/June, 2017.

[17] H. Ballani, K. Jang, T. Karagiannis, C. Kim, D. Gunawardena, G. O'Shea, Chatty Tenants and the Cloud Network Sharing Problem, *10th USENIX Symposium on Networked Systems Design and Implementation (USENIX NSDI' 13)*, Lombard, IL, 2013, pp. 171-184.

[18] W. Li, D. Guo, K. Li, H. Qi, J. Zhang, Idaas: Inter-datacenter Network as a Service, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 29, No. 7, pp. 1515-1529, July, 2018.

[19] X.-C. Sun, Z.-J. Wang, H. Chu, Q.-R. Zhang, An Efficient Resource Management Algorithm for Information Centric Networks, *Journal of Internet Technology*, Vol. 17, No. 5, pp. 1007-1015, September, 2016.

[20] X. Liu, P. Wang, Z. Lan, B. Shao, Biological Characteristic Online Identification Technique over 5G Network, *IEEE Wireless Communications*, Vol. 22, No. 6, pp. 84-90, December, 2015.

Biographies

Xiaocui Sun received the Ph.D. degree in the Department of Computing at the Hong Kong Polytechnic University in 2011. She is currently working in the Department of Computer Science at Guangdong Pharmaceutical University. Her research interest includes Network resource management, Traffic engineering, Medical Information, Internet Architecture and Protocols.

Xinran Zhuo is with the Glasgow College, University of Electronic Science and Technology of China. Her research interests are multimedia signal processing, high-speed network, network security and data management in mobile networks.

Zhijun Wang is current a visiting scholar at The University of Texas at Arlington. His research is focusing on traffic control, job scheduling and resource management in cloud computing and edge computing.