This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2015.2497701, IEEE Transactions on Parallel and Distributed Systems

New Bandwidth Sharing and Pricing Policies to Achieve a Win-Win Situation for Cloud Provider and Tenants

Haiying Shen, *Senior Member, IEEE,* and Zhuozhao Li

Abstract—For predictable application performance or fairness in network sharing in clouds, many bandwidth allocation policies have been proposed. However, with these policies, tenants are not incentivized to use idle bandwidth or prevent link congestion, and may even take advantage of the policies to gain unfair bandwidth allocation. Increasing network utilization while avoiding congestion not only benefits cloud provider but also the tenants by improving application performance. In this paper, we propose a new pricing model that sets different unit prices for reserved bandwidth, the bandwidth on congested links and on uncongested links, and makes the unit price for congested links proportional to their congestion degrees. We use game theory model to analyze tenants' behaviors in our model and the current pricing models, which shows the effectiveness of our model in providing the incentives. With the pricing model, we propose a network sharing policy to achieve both min-guarantee and proportionality, while prevent tenants from earning unfair bandwidth. We further propose methods for each virtual machine to arrange its traffic to reduce its unsatisfied demand and maximize its utility, while increase network utilization. As a result, our solution creates a win-win situation, where tenants strive to increase their benefits in bandwidth sharing, which also concurrently increases the utilities of cloud provider and other tenants. Our simulation and trace-driven experimental results show the effectiveness of our solutions in creating the win-win situation.

Index Terms—bandwidth allocation, pricing policies, network proportionality, min-guarantee

1 Introduction

Cloud computing attracts many enterprises (e.g., Dropbox, Facebook video storage) to migrate their business or services to the clouds without the need to build their own datacenters. Cloud provider (provider in short) multiplexes computation, storage and network resources among different tenants, enabling them to independently run their own jobs on the cloud. Nowadays, on the Infrastructure as a Service (IaaS) (e.g., Amazon EC2), the resources are charged based on the renting time period of virtual machines (VMs) and VM types (with different CPU and memory storage) [1]. Though the CPU and memory storage of a VM are dedicated resources to a tenant, each network link is shared among tenants, which makes it non-trivial to guarantee the provision of a certain bandwidth to a tenant. Current best-effort bandwidth provision is insufficient to guarantee the quality-of-service to tenants (i.e., satisfy Service Level Objective (SLO)). Congested links lead to slow traffic rate, which not only degrades the performance of tenants' applications but also increases their cost due to longer VM usage.

Previous research studied the problem of bandwidth-Previous research studied the problem of bandwidth allocation among tenants with different requirements (e.g., minimum guarantee [2], [3] , high utilization, and network proportionality [3]–[10]). *Min-guarantee* means guaranteeing the minimum bandwidth that tenants expect for each VM, irrespective of the network utilization of other tenants. *High utilization* means maximizing network utilization in the presence of unsatisfied demands. This means an application can use the idle bandwidth, which shortens job completion time (that benefits tenants) and enables more jobs to be deployed in the infrastructure (that increases the provider's revenue). *Network proportionality* means that network resources allocated to tenants are proportional to their payments, which aims to achieve fairness between tenants. For example, proportionality is considered in allocating bandwidth in a congested link when several tenants compete bandwidth on the link in order to achieve fairness. Popa *et al.* [3] indicated that a desirable allocation solution should meet three requirements: *min-guarantee*, *high utilization* and *network proportionality*, which however are difficult to achieve simultaneously due to their tradeoffs.

The tradeoff between min-guarantee and network proportionality means that the min-guarantee demands of VMs with low payments cannot be satisfied due to the domination of VMs with much higher payments. The tradeoff between high utilization and network proportionality means that with the bandwidth allocation policies for network proportionality, tenants are not incentivized to use uncongested links or even are incentivized to reduce actual bandwidth demand on some links, which reduces network utilization. For example, a tenant tries to compete bandwidth in a more-important congested link even though it can use an idle link; it may also purposely change its actual bandwidth demand to receive more bandwidth allocation [3]. There are two essential reasons for the tradeoffs. First, tenants would try to gain more benefits at the cost of the provider or other tenants. Second, bandwidth allocation based on predetermined proportionality enables tenants to take advantage of it to gain more allocated bandwidth. The tradeoff can be resolve if we can achieve a win-win situation, in which when tenants strive to increase their utility in bandwidth sharing, it also concurrently increases the network utilization, profit and SLO conformance of the provider.

To achieve the win-win situation, we propose a new bandwidth pricing model for bandwidth allocation. Our pricing model considers three parts in determining the

• H. Shen and Z. Li are with Department of Electrical and Computer Engineering Clemson University, Clemson, SC 29631.
Email: {shenh, zhuozhl}@clemson.edu

payment of a tenant (P_{t_i}): min-guarantee bandwidth (M_{t_i}), consumed bandwidth on congested links ($B^c_{t_i}$) and on uncongested links ($B^u_{t_i}$) of all VMs of the tenant. That is, $P_{t_i} = (\alpha M_{t_i} + \beta B^c_{t_i} + \gamma B^u_{t_i})/2$ ($\alpha > \beta > \gamma$), where α, β and γ are unit prices and β is proportional to link congestion degree. Therefore, to reduce payment, a tenant will buy the minimum bandwidth on a VM based on its real minimum demand (i.e., min-guarantee), which reduces the provider's reserved but unused resources and increases network utilization. Also, a tenant will try to use idle bandwidth and avoid more congested bandwidth, which increases network utilization and decreases unsatisfied bandwidth demands. High network utilization in turn increases the performance of applications and hence benefits tenants and the provider.

With this pricing model, the two aforementioned essential reasons can be avoided. First, rather than using flat-rate per VM payment model, which stimulates tenants to compete for bandwidth since the consumed bandwidth does not affect payment, our pricing model determines the payment based on actually consumed bandwidth and sets different prices for min-guarantee bandwidth and consumed bandwidth on uncongested and congested links, which encourages tenants to be cooperative (e.g., avoiding congested links, using uncongested links, limiting min-guarantee) in bandwidth sharing to reduce their payment. Second, rather than allocating bandwidth based on tenant payment on purchased VMs (i.e., payment-first-allocation-second), our pricing model determines each tenant's payment based on allocated bandwidth after bandwidth allocation (i.e., allocation-first-payment-second). Thus, the allocated bandwidths of tenants are always proportional to their payments for each of the three types of bandwidths, which achieves network proportionality to a certain extent.

We also propose several bandwidth allocation policies that resolve the tradeoff between min-guarantee and network proportionality to a certain extent, and strengthen the win-win situation, i.e., satisfying tenant demands as much as possible while increasing network utilization. We first satisfy the min-guarantee, and then achieve proportionality (network, congestion or link proportionality [3]) on the residual bandwidth. With our pricing model, tenants are disincentivized to take advantage of the allocation policies (or even cheat) for more bandwidth, which increases network utilization [3]. As a result, our solution creates the win-win situation, which leads to less unsatisfied bandwidth demands, higher network utilization and fewer congested links that benefit not only the tenants but also the provider. It also helps simultaneously achieve the above-stated three requirements – an unsolved problem in previous research within our knowledge.

Below, we summarize the contributions of our paper:

- We use the game theory model to analyze the behaviors of tenants in the current pricing models and allocation policies. We find that tenants may try to gain more benefits at the cost of the provider and other tenants.
- We propose a pricing model to create a win-win situation, where tenants try to gain more utility which also concurrently increases the benefits of other tenants and the provider. Our analysis on the tenant behaviors confirms the advantages of our pricing model.
- We propose a network sharing policy to achieve both min-guarantee and different types of proportionality, while preventing tenants from earning unfair bandwidth.
- We propose a foreign link transmission policy that encourages a VM to transmit its traffic through multiple least congested links when the least congested link cannot satisfy its demand, which reduces unsatisfied tenant demands and increases network utilization.
- We propose a bandwidth allocation enhancement policy that transfers VMs' extra allocated bandwidth beyond their demands to VMs with unsatisfied demands, which reduces unsatisfied demands and increases network utilization.
- We propose a traffic flow arrangement policy for each VM to determine the links to traverse its traffic flows to their destinations, and the destination VMs for flows without fixed destinations in order to maximize the number of its satisfied demands while increasing network utilization.

Consequently, with our solution, the competitive cloud environment is transformed to a cooperative environment, which increases the benefits of both the provider and tenants, and helps create a harmonious ecosystem. Our experimental results verify the advantages of our solution. The rest of this paper is structured as follows. Section 2 presents a review of related work. Section 3 analyzes the behaviors of tenants in current bandwidth allocation and pricing model and shows that competitive bandwidth sharing does not benefit either tenants or the provider. Section 4 presents our proposed policies, and analyzes their effectiveness in increasing the benefits of both sides. Section 5 presents the performance of our proposed policies in comparison to previous bandwidth allocation strategies. Finally, Section 6 concludes this paper with remarks on our future work.

2 Related Work

Recently, many bandwidth allocation mechanisms have been proposed. Some works [2], [3] provide proportional network sharing based on VM weight (or payment), while other works [3]–[10] provide minimum guarantee.

Seawall [2] is a hypervisor-based mechanism to enforce the bandwidth allocation in each congested link based on the weights of the VMs which are communicating along that link. Popa *et al.* [3] proposed PS-L and PS-N to achieve proportionality. PS-L achieves *link proportionality*, in which the allocated bandwidth in a congested link is proportional to the sum of the weights of a tenant's VMs that communicate through the link. PS-N achieves *congestion proportionality*, in which the total allocated bandwidth on congested links of a tenant is proportional to the sum of the weights of a tenant's VMs. Although these policies can achieve proportionality, they cannot provide min-guarantee for predictable performance.

Popa *et al.* [3] also proposed PS-P to support minimum bandwidth guarantees by assigning the weight of on link between a VM-pair based on the weight of the VM closer to the link. Oktopus [4] and SecondNet [5] use static reservations in the network to achieve minimum bandwidth guarantees. Guo *et al.* [6] proposed to achieve min-guarantee and then share the residual bandwidth among VM-pairs for link proportionality. ElasticSwitch [7] utilizes the spare bandwidth from unreserved capacity or underutilized reservation to provide bandwidth guarantee and achieve high utilization. EyeQ [8] is a system that leverages the high bisection bandwidth of datacenter and enforces admission control on traffic in order to guarantee the minimum bandwidth to the tenants. Cicada [9] is a system to predict tenants' bandwidth guarantee requests by a weighted linear combination of previous observations, which helps improve network utilization. Lee *et al.* [10] presented CloudMirror, which provides bandwidth guarantee for cloud applications based on a new network abstraction – Tenant Application Graph (TAG) rather than the hose model, and a VM placement strategy to efficiently utilize the resources. However, the above works do not provide network proportionality.

In the above works, since bandwidth is allocated based on weight determined by flat-rate payment, all tenants try to compete for bandwidth, which reduces network utilization and increases SLO violations. Different from these policies, our solution provides utilization incentives, and simultaneously achieves the three aforementioned requirements.

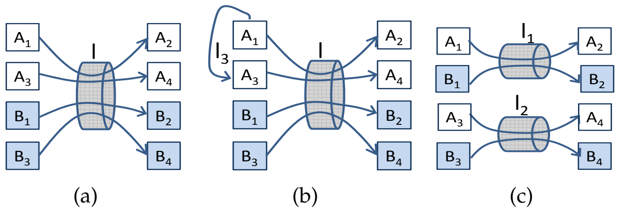
Niu *et al.* [11] proposed a pricing model for cloud bandwidth reservation to maximize social welfare. Feng *et al.* [12] utilized the bargaining game to maximize the resource utilization in video streaming datacenters. Wilson *et al.* [13] proposed a congestion control protocol to allocate bandwidth according to flow deadlines, and charge bandwidth usage. Guo *et al.* [14] proposed a bandwidth allocation policy via a cooperative game approach to achieve minimum bandwidth guarantee and fairness among VMs. In order to handle the dynamic nature of datacenter traffic, the authors [15] presented a distributed bandwidth algorithm, which not only provides minimum guarantee but also provides fast convergence to fairness and smooth response to bursty traffic. However, their works did not consider incentivizing the tenants to use the uncongested links for high network utilization. Different from these pricing models, our pricing model aims to provide incentives to tenants in order to use uncongested links to increase network utilization, and prevent congestion to reduce SLO violations, which creates a win-win situation for both the provider and tenants.

Bandwidth pricing in broadband networks has been studied previously. Kelly [16] addressed the issues of charging, rate control and routing for elastic traffic. A user's allocated bandwidth is determined by how much the user is willing to pay. Kelly *et al.* [17] further proposed a solution to the network utility maximization problem in order to allocate network resources in a fair and distributed manner to the users in a large-scale network. They also allow routing control, which may be naturally implemented with proportionally fair pricing. MacKie-Mason *et al.* [18] indicated that usage-based pricing is necessary for congestion control on Internet and proposed per-packet prices that vary according to the degree of congestion. These works lay the foundation for the design of our pricing policies in the cloud.

3 Competitive Bandwidth Sharing in Current Policies

3.1 Problems in Bandwidth Allocation and Our Solutions

In this section, we first explain the two tradeoffs indicated in Section 1 in detail in order to show that when tenants try to gain more bandwidth, the network utilization may be decreased. We then show that with the current pricing models, the pursuit of higher utility of a tenant may decrease the utility of the other tenants and the provider. Therefore, the goal of our work is to create a win-win situation, where all tenants cooperate to increase their utilities and also concurrently increase the system utility, which not only benefits all tenants but also the provider. To achieve this goal, we propose our pricing model and network sharing policy in the next section.



(a) (b) (c)

Fig. 1: Failure to achieve high utilization in bandwidth allocation [3].

When developing a bandwidth allocation policy, rather than aiming to meet a part of the three requirements (i.e., min-guarantee, high utilization and network proportionality), we should simultaneously meet the three requirements to achieve the ultimate goal, i.e., increasing the profit of the provider and the performance of tenants' applications based on their payments. Below, we present the unsolved problems in bandwidth allocation indicated in [3] that prevent us from simultaneously achieving these requirements and briefly explain our solutions.

3.1.1 *Tradeoff Between Min-guarantee and Network Proportionality*

Suppose tenant A employs 2 VMs and tenant B employs 10 VMs. We assume the weights of VMs are the same for simplicity. VMs A_1 and B_1 are hosted on the same physical machine (PM) that communicate with other VMs that belong to the same tenant. According to the network proportionality, A_1 receives $2/12$ of the access link, while B_1 receives $10/12$. A_1's allocation may be lower than its minimum guarantee, failing to satisfy its min-guarantee. Also, tenant B can buy many VMs for B_1 to communicate in order to dominate the link, which would degrade A's application performance. To address this tradeoff problem, we first satisfy the min-guarantee of each VM and then follow the network proportionality in allocating the residual bandwidth. We also set the highest unit price for the min-guarantee bandwidth, so that tenants will try to limit the minimum bandwidth to their exact needs, which prevents the domination situation to a certain degree.

3.1.2 *Tradeoff Between High Utilization and Network Proportionality*

Consider two tenants A and B, each employing 4 VMs with the same weight. Their flows traverse the same congested link l with capacity C as shown in Figure 1(a). Based on the network proportionality, each tenant receives $C/2$ bandwidth. Now assume VMs A_1 and A_3 start communicating along an uncongested path l_3 (Figure 1(b)). In order to maintain network proportionality, tenant A's allocation is decreased along link l. If A's traffic along l is more important than that along path l_3, A is disincentivized to use path l_3, which degrades network utilization and also increases the probability of link congestion.

To address this problem, we assign lower unit price to uncongested links than congested links and make the unit price for congested links proportional to the congestion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2015.2497701, IEEE Transactions on Parallel and Distributed Systems

degree. Then, tenants are incentivized to use uncongested links, and avoid competing for bandwidth in the congested links. The higher the congestion of a link, the lower probability for a tenant to compete for bandwidth on the link. As a result, the network utilization is increased and the congestions are prevented or mitigated, which enhances application performance and also increases the provider's profit and reduces SLO violations.

Popa *et al.* [3] indicated that congestion proportionality can achieve utilization incentives but tenants may cheat to gain more bandwidth which reduces network utilization. Since the uncongested links are not considered in bandwidth allocation, tenants are incentivized to use uncongested links. However, a tenant can reduce its demand on purpose to change a congested link to an uncongested link in order to increase its own allocation and reduce others' allocation, which decreases network utilization. Assume ϵ is a very small number. In Figure 1(c), if the demand of $B_3 \rightarrow B_4 = \epsilon$, the allocation $A_3 \rightarrow A_4 = C - \epsilon$, and then $B_1 \rightarrow B_2 = C - \epsilon$ and $A_1 \rightarrow A_2 = \epsilon$. Tenant A can purposely change its demands on l_2 to $C - 2\epsilon$. Then, l_2 becomes uncongested and is not considered in congestion proportionality. Finally, tenant A receives $3C/2 - 2\epsilon$ and tenant B receives $C/2 + \epsilon$. The network utilization is decreased from $2C$ to $2C - \epsilon$.

Suppose D_l and C_l denote the total bandwidth demand and capacity on link l, we argue that *congested links* should be defined as the links with $D_l > C_l$ rather than $D_l \geq C_l$ as in [3] and *uncongested links* should be defined as the links with $D_l \leq C_l$. Because when $D_l = C_l$, the link can exactly satisfy the tenants' demands and there is no need for them to compete for bandwidth. With this new definition, a tenant only has incentives to purposely reduce its demand when $D_l > C_l$ to make it $D_l = C_l$ (the link is fully utilized), in which the tenants have no incentives to reduce their demands. In a congested link, each tenant checks its gain and cost to decide if it should reduce demand to make it $D_l = C_l$. The gain includes more allocation in other congested links and lower payment in our pricing model. Note that instead of preventing tenants from reducing their demands when $D_l > C_l$, we encourage such behavior, because it will not reduce network utilization and avoids link congestion, which increases application performance for tenants and reduces SLO violations of the provider. Though finally tenant A may receive more bandwidth in another congested link, it still needs to pay for this bandwidth in our pricing model, which achieves proportionality.

3.2 Game Theory Based Analysis on Current Pricing Models

We analyze the behaviors of tenants and the provider using the non-cooperative game theory [19], in which each player tries to maximize its payoff. We first analyze the current price model in Amazon EC2, where tenants pay a fixed flat-rate per VM for each type of VMs. When a link is congested, a previously proposed bandwidth allocation strategy (min-guarantee, network proportional, congestion proportionality or link proportionality) is used. Currently, the provider supplies bandwidth in the best-effort provision manner. Therefore, we assume that without the min-guarantee requirement, the bandwidth provision does not affect the SLO violations, and with this requirement, failures of providing the min-guarantee bandwidth lead to SLO violations.

The utility of the provider (i.e., cloud profit) is the difference between its total revenue and total cost, which includes the cost for consumed bandwidth and for SLO violations. We use N_{V_i} $(1 \leq i \leq m)$ to represent the total number of sold type-i VMs, use m to represent the number of VM types in the system and use p_i to denote the payment of a type-i VM. We use b to denote the unit cost to the provider for the usage of each bandwidth unit due to power consumption, hardware wear and tear and etc. B^a denotes the allocated bandwidth of all tenants and $B^a_{t_i}$ denotes the allocated bandwidth of tenant t_i. M_{v_i} denotes the min-guarantee bandwidth for VM v_i. The min-guarantee bandwidth for tenant t_i (M_{t_i}) is the sum of the min-guarantee bandwidths of t_i's VMs: $M_{t_i} = \sum_{v_k} M_{t_i,v_k}$. We use $H_{t_i} = M_{t_i} - B^a_{t_i}$ to denote the unsatisfied bandwidth for t_i to meet the min-guarantee requirement. It leads to $F_c(H_{t_i})$ utility loss of the provider caused by the reputation degradation and potential revenue loss. We use $F_{t_i}(H_{t_i})$ to denote the utility loss of tenant t_i due to unfilled demands from clients. With the min-guarantee requirement, reserving bandwidth capacity M_{t_i} will incur a reservation cost of cM_{t_i} [11]. Then, the provider's utility can be represented by:

$$U_c = \begin{cases} \sum\limits_i p_i N_{V_i} - bB^a, & \text{w/o min-g} \\ \sum\limits_i p_i N_{V_i} - bB^a - \sum\limits_{t_i} F_c(H_{t_i}) - cM_{t_i}, & \text{w/ min-g,} \end{cases} \quad (1)$$

in which "min-g" denotes min-guarantee requirement. A tenant's utility can be represented by:

$$U_{t_i} = g_{t_i} B^a_{t_i} - \sum_k p_k N_{V_{k\,t_i}} - F_{t_i}(H_{t_i}), \quad (2)$$

where g_{t_i} represents the earned utility of each used bandwidth unit and $N_{V_{k\,t_i}}$ denotes the number of type-k VMs bought by tenant t_i.

Based on Equation (1), for the provider, in order to maximize its utility, it needs to increase the number of sold VMs (N_{V_i}), and reduce the total used bandwidth (B^a). With min-guarantee, the provider also needs to reduce provision failure on reserved bandwidth (reduce congestion) and reduce reserved bandwidth. Given a certain number of PMs, to increase N_{V_i}, the provider can place many VMs on one PM. To reduce B^a, the provider can employ strategies such as placing the VMs of the same tenant in the same or nearby PMs (which is out of the scope of this paper). Given a certain VM placement, the provider supplies bandwidth in the best-effort manner, and it has no control over B^a. Consequently, it tries to maximize the number of VMs placed in a PM while guarantee the minimum bandwidth for VM and reduce link congestion. Though the provider can use bandwidth allocation policies to achieve different proportionality, it has no control on tenants' bandwidth demands to reduce the link congestion situation. Thus, the provider needs an additional policy for this purpose to increase cloud profit.

Based on Equation (2), in order to increase utility, a tenant tries to receive more $B^a_{t_i}$, buy fewer and less-expensive VMs and also reduce its unsatisfied demand. As a result, tenants will try to be economical when buying VMs and compete for more bandwidth. As explained in Section 3.1, in the network proportionality or congestion proportionality policy, the competition leads to low network utilization, which reduces the utility of the provider and other tenants.

We then analyze the recently proposed pricing model in [11]. Each tenant pays p for every unit bandwidth consumed and pays $k_{t_i}w_{t_i}$ for having w_{t_i} portion of its demand guaranteed. Then, the utilities of provider and tenant are:

$$U_c = \sum_{t_i}(pB^a_{t_i} + k_{t_i}w_{t_i}) - \sum_{t_i} F_c(w_{t_i}D_{t_i} - B^a_{t_i}) - cM_{t_i}, \quad (3)$$

$$U_{t_i} = g_{t_i}B^a_{t_i} - (pB^a_{t_i} + k_{t_i}w_{t_i}) - F_{t_i}(w_{t_i}D_{t_i} - B^a_{t_i}). \quad (4)$$

Equation (3) indicates that to increase utility, the provider wishes to increase network utilization (B^a) and reduce unsatisfied demands. However, it has no control on bandwidth demands from tenants. Equation (4) shows that to maximize its utility, given a reserved portion, a tenant tends to compete for bandwidth in demand. Since the unit price for used bandwidth is the same regardless of the congestion degree of links, tenants tend to compete for more important bandwidth to them, as shown in Figure 1(b).

Both pricing models lead to bandwidth competition among tenants. As explained in Section 3.1, though different allocation policies can be used in bandwidth competition, the competition still can lead to low network utilization and reduce the benefits of other tenants and the provider. That is, the pursuit of higher utility of a tenant decreases the utility of the other tenants and the provider. We need a policy to create a harmonious environment where all tenants cooperate to increase their utilities and also concurrently increase the system utility and reduce unsatisfied demands, which not only benefits all tenants but also the provider.

4 Proposed Policies for Cooperative Bandwidth Sharing

In this section, we present our per VM-pair pricing model that can achieve high network utilization and also avoid congested links, thus increase application performance and reduce SLO violations. More importantly, this pricing model transforms the competitive environment to a cooperative environment, in which a tenant can receive more benefits by being cooperative than by being non-cooperative.

Current datacenter network topologies are generally fat-tree topologies [20]–[24]. Therefore, in this paper, we assume a multi-path or multi-tree topology, where each VM has multiple links to connect to other VMs, as shown in Figure 2. We only drew the multiple links for A_1 and A_9 as an example for easy readability. Typically there are several layers of switches in datacenter networks. While VM communications can have full bisection bandwidth within a rack, modern production clusters typically have oversubscription for the bandwidth between the ToR and core switches [4], [22]. Hence, it is more reasonable to charge the more competitive cross-rack bandwidth rather than within-rack bandwidth. In addition, the traffic from a VM on the links between the ToR and aggregate switches may not pass the links between the aggregate and core switches. Then, charging the bandwidth between the aggregate and core switches may result in some used bandwidth of a VM uncharged. Therefore, in this paper, we consider the bandwidth allocation and pricing on the links between the ToR and aggregate switches. In our pricing model, a VM is only charged for its consumed bandwidth on these links.

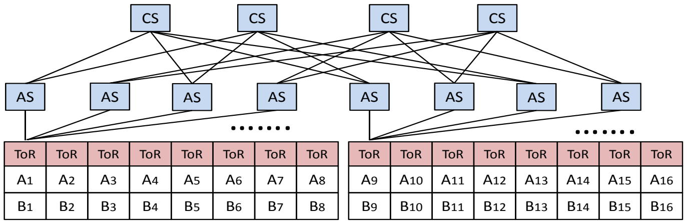


Fig. 2: An example of multi-tree topology [22]. ToR: top-of-rack switch; AS: aggregate switch; CS: core switch.

4.1 A New Bandwidth Pricing Model

When a tenant buys VMs, it can specify the min-guarantee of each VM though it does not have to. If a tenant demands a minimum bandwidth for its VM, the cloud provider reserves the requested min-guarantee for the VM and charges this reserved resource based on the price. Previous works [3], [4], [10], [25] indicate that the VMs' traffic demands can be predicted. However, if a tenant cannot do the prediction, it can choose not to demand minimum bandwidths. The pricing model used by the cloud provider only handles the minimum bandwidth demands from the tenants and does not concern about the accuracy of the demand prediction.

We use *congested bandwidth* ($B^c_{t_i}$) and *uncongested bandwidth* ($B^u_{t_i}$) to represent tenant t_i's consumed bandwidth on congested links and on uncongested links, respectively. Then, t_i's total allocated bandwidth $B^a_{t_i} = B^u_{t_i} + B^c_{t_i}$. We use M_{t_i,v_j}, $B^c_{t_i,v_j}$ and $B^u_{t_i,v_j}$ to represent the minimum guaranteed bandwidth, the congested and uncongested bandwidth of VM v_j of tenant t_i; $B^a_{t_i,v_j} = B^u_{t_i,v_j} + B^c_{t_i,v_j}$. We use α, β and γ to denote the unit price of minimum guaranteed bandwidth, congested bandwidth and uncongested bandwidth and $\alpha > \beta > \gamma$. Then, each tenant's payment consists of three parts:

$$\begin{aligned} P_{t_i} &= (\alpha M_{t_i} + \beta B^c_{t_i} + \gamma B^u_{t_i})/2 \\ &= (\alpha \sum_{v_j} M_{t_i,v_j} + \beta \sum_{v_j} B^c_{t_i,v_j} + \gamma \sum_{v_j} B^u_{t_i,v_j})/2 \quad (5) \end{aligned}$$

Since the pricing model is based on per VM-pair, coefficient $\frac{1}{2}$ is used to avoid charging the used bandwidth on each link twice. For tenants, the reserved bandwidth is more valuable than non-reserved bandwidth, because a tenant is guaranteed to receive the reserved bandwidth. Therefore, it should pay more for reserved bandwidth. If its price is low, each tenant would try to buy more minimum bandwidth, which would generate much reserved but unused bandwidths and hence reduce the cloud profit. Reserved bandwidth (M_{t_i}) incurs additional cost of cM_{t_i} to the provider. On the other hand, it reduces the utility loss due to poor performance of applications. Then, to increase profit, the provider should encourage tenants to reserve no more bandwidth than their exact needs, which also increases network utilization. Thus, we set α to the highest value among the unit prices, i.e., $\alpha > \beta, \gamma$. With this pricing policy, the tenants have no incentives to lie for their min-guarantee demands, since they need to pay for their min-guarantee, which has much higher unit price than the normally consumed bandwidth. With this pricing policy, tenants will be incentivized to more accurately estimate their exact usage as the min-guarantee bandwidth since overestimation will cost them more and underestimation will degrade the performance of their applications. Even if they report untruthful information for their predicted amount, it will not impair the benefits of the cloud provider since it earns much profit from the reserved bandwidth.

In the ideal situation, each link achieves $D_l = C_l$; i.e., the network is fully utilized and all bandwidth demands are satisfied. Then, both the provider and tenants earn the maximum profit and experience the least utility loss due to unfulfilled demands. To make the system approach the ideal situation, we need to encourage tenants to use uncongested links and avoid using congested links. Therefore, the unit price (β) of congested bandwidth should be higher than the unit price (γ) of uncongested bandwidth. To tenants, congested bandwidth is more valuable than uncongested bandwidth as they must compete for it. With $\beta > \gamma$, tenants are incentivized to use uncongested links and avoid using congested links to reduce payment.

We define a link's *congestion degree* as $\frac{D_l}{C_l}$. To avoid exacerbating the congestion situation, the tenants should be more strongly disincentivized to use more congested links. Thus, we set a congested link's β to be proportional to its congestion degree: $\beta = \gamma(\min\{\frac{D_l}{C_l}, \delta\})$ ($\frac{D_l}{C_l} > 1$). $\delta > 1$ is used to limit the infinite increase of β.

4.2 Network Bandwidth Sharing

To consider both min-guarantee and proportionality in a congested link, each VM first receives its min-guarantee, and then receives its share on the residual bandwidth based on the proportionality allocation policy, which can be network proportionality, congestion proportionality or link proportionality. Let D_{v_i} denote the total demand of VM v_i. Then, we have $D_{v_i} = \sum_{v_k} D_{v_i,v_k}$, where v_k denotes each VM that v_i communicates with and D_{v_i,v_k} denotes the traffic demand between VM v_i and v_k. The total bandwidth allocated to VM v_i equals $B^a_{v_i} = \sum_{v_k} B^a_{v_i,v_k}$. Below, we first introduce a method to calculate the min-guarantee bandwidth for a pair of VMs to ensure that the min-guarantee of each VM is guaranteed. Then, we introduce how to calculate the weight of a pair of VMs. Finally, we introduce the entire process of bandwidth requesting and allocation.

VM v_i may communicate with many other VMs through a link, as shown in Figure 3. Therefore, to ensure that $B^a_{v_i}$ satisfies M_{v_i}, v_i's min-guarantee should be distributed among these VMs. Assume that v_j is one of these communicating VMs. Hence, v_j should receive its portion equals to M_{v_j} over the sum of the min-guarantee of all these VMs, i.e., $M_{v_i} \frac{M_{v_j}}{\sum\limits_{D_{v_i,v_k} \neq 0} M_{v_k}}$. Similarity, v_i should receive $M_{v_j} \frac{M_{v_i}}{\sum\limits_{D_{v_j,v_k} \neq 0} M_{v_k}}$. We use a similar method as in [3] to determine M_{v_i,v_j}. That is, we define the min-guarantee of a pair of VM v_i and v_j over a link as:

$$M_{v_i,v_j} = \rho M_{v_i} \frac{M_{v_j}}{\sum\limits_{D_{v_i,v_k} \neq 0} M_{v_k}} + (1-\rho) M_{v_j} \frac{M_{v_i}}{\sum\limits_{D_{v_j,v_k} \neq 0} M_{v_k}}, \quad (6)$$

where $\rho = 1$ for all links in the tree topology that are closer to v_i than v_j, and $\rho = 0$ for all links closer to v_j than v_i. Equation (6) calculates the minimum bandwidth of a VM pair over a link based on which node is closer to the link, i.e., is more likely to use this link. For example, in Figure 2, if VM A_1 communicates with A_9 through a link that connects A_1's ToR switch and an aggregate switch, since A_1 is closer to the link, $\rho = 1$ and M_{v_i,v_j} is calculated based on the first part of Equation (6). After the calculation of the minimum bandwidth of a VM pair over a link, the cloud provider guarantees to provide this minimum bandwidth during the VM pair's communication on this link. In this way, the minimum bandwidth of v_i over the link is guaranteed to be M_{v_i}, since the sum of the minimum bandwidths for v_i's communication pairs on this link is M_{v_i}. Hence, we achieve min-guarantee for VMs on a link.

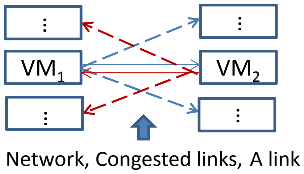


Fig. 3: Communication between VMs.

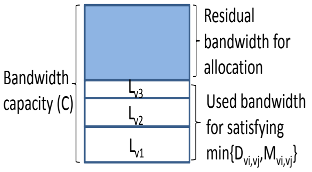

Residual bandwidth for allocation
Bandwidth capacity (C)
L$_{v3}$
L$_{v2}$
L$_{v1}$
Used bandwidth for satisfying min{D$_{vi,vj}$,M$_{vi,vj}$}

Fig. 4: Bandwidth sharing in a link.

Suppose VM v_i demands bandwidth D_{v_i,v_j} to VM v_j on a link. We define: $L_{v_i,v_j} = \min\{D_{v_i,v_j}, M_{v_i,v_j}\}$. If the link has residual bandwidth no less than L_{v_i,v_j}, v_i receives $L_{i,j}$ and there is no competition on the link. Otherwise, each pair of communicating VMs $v_{i'}$ and $v_{j'}$ on the link receive their $L_{i',j'}$, and then the residual bandwidth is allocated among the pair of VMs that have unsatisfied demands based on proportionality.

We directly use the min-guarantees of VMs as the weights of VMs in bandwidth allocation. The cloud can also specify different levels of competition ability for the tenants to purchase as the weights of VMs in bandwidth competition. The weight of a pair of VMs v_i and v_j on a link equals:

$$W_{v_i,v_j} = M_{v_i} \frac{M_{v_j}}{\sum\limits_{D_{v_i,v_k} \neq 0} M_{v_k}} + M_{v_j} \frac{M_{v_i}}{\sum\limits_{D_{v_j,v_k} \neq 0} M_{v_k}}, \quad (7)$$

As shown in Figure 3, $\sum_{D_{v_i,v_k} \neq 0} M_{v_k}$ means the sum of the min-guarantees of all VMs that v_i communicates with through this link, across the entire network, and in all congested links in the link proportionality, network proportionality and congestion proportionality policy, respectively.

In the following, we explain the process of bandwidth requesting and allocation with our pricing model. The implementation of the policy can rely on switch support or hypervisors as explained in [3]. We can use rate-limiters at hypervisors to throttle the bandwidth, and the switch support can help the VMs to pick appropriate paths based on the policies. When VM v_i declares its bandwidth demand to v_j on a link, if the residual bandwidth is no less than the demand, v_i receives its demanded bandwidth. Otherwise, the link will be congested and the unit price for the bandwidth on this link increases. In this case, v_i can consider if it can reduce its demand to make $D_l = C_l$ based on the traffic's delay tolerance. Recall we assume a multi-path or multi-tree topology. v_i can also seek other alternative uncongested links. If it must make a demand that leads to $D_l > C_l$, the VMs on the link are notified the possible congestion. Since the congestion leads to higher unit price for all VMs on the link, the VMs will try to constrain the link congestion degree. Since some applications are delay-tolerant (e.g., high-throughput computing task) while others are delay-sensitive (e.g., VoD applications), the VMs of delay-tolerant applications can reduce their bandwidth demands if its performance degradation is tolerable. We will explain how a notified VM can decide the amount of bandwidth reduction in Section 4.3. The notified VMs also seek other alternative uncongested links to transmit data. Then, if the link still will become congested, as shown in Figure 4, the L_{v_i,v_j} of each VM should be first satisfied, and the residual bandwidth will be allocated among VMs with $B^a_{v_k} < D_{v_k}$ using our allocation policy. Since higher congestion links have higher unit price, VMs are incentivized to reduce demands hence congestion in order to reduce their payments. As each tenant tries to avoid congested links and use uncongested links, and also constrain the congestion degrees of congested links, the network utilization is increased and the SLO violations are reduced, which benefits the provider and also the tenants. Note that in our bandwidth sharing policy, the cloud provider provides a tenant the available bandwidth capacity information of the alternative links connecting to the tenant's VM between the ToR switches and aggregate switches. All information learned by the tenant only includes several available bandwidth values and it only needs to indicate that it selects the link identified by its available capacity. Therefore, the tenants do not know any other information of the topology and the topology remains secure and private.

4.3 Analysis of Our Pricing Model

We use R_c to denote the provider's revenue. In our pricing model, the utility of the provider equals:

$$\begin{aligned} U_c &= R_c - bB^a - \sum_{t_i} F_c(H_{t_i}) - cM_{t_i} \\ &= \sum_{t_i}\{(\alpha M_{t_i} + \beta B^c_{t_i} + \gamma B^u_{t_i}) - b(B^c_{t_i} + B^u_{t_i}) - F_c(H_{t_i}) - cM_{t_i}\} \\ &\geq (\alpha - c)\sum_{t_i} M_{t_i} + (\gamma - b)B^a - \sum_{t_i} F_c(H_{t_i}) \end{aligned} \tag{8}$$

G_{t_i} denotes the gain of tenant t_i from receiving bandwidth, P_i denotes the payment and O_i denotes the utility loss due to unsatisfied demand. The utility of a tenant equals:

$$\begin{aligned} U_{t_i} &= G_{t_i} - P_{t_i} - O_{t_i} \\ &= g_{t_i}(B^c_{t_i} + B^u_{t_i}) - (\alpha M_{t_i} + \beta B^c_{t_i} + \gamma B^u_{t_i}) - F_{t_i}(H_{t_i}) \end{aligned} \tag{9}$$

$(g_{t_i}-\beta)B^a_{t_i}-\alpha M_{t_i}-F_{t_i}(H_{t_i}) \leq U_{t_i} \leq (g_{t_i}-\gamma)B^a_{t_i}-\alpha M_{t_i}-F_{t_i}(H_{t_i})$. Based on Equation (8), for the provider, in order to increase utility, it needs to increase B^a, i.e., increase the network utilization, sell more reserved bandwidth, and decrease unsatisfied demands.

As indicated in Section 3.2, in current pricing models, the provider has no control on how much and in which links tenants demand bandwidth. Only when a link is congested, the provider allocates the bandwidth among tenants based on min-guarantee or proportionality. Therefore, the provider cannot actively try to increase its utility. Using our proposed pricing model, the provider can guide how much and in which links that tenants demand bandwidth to increase their utility, which in turn increases the provider's utility.

Equation (9) shows that in order to increase utility, a tenant needs to gain more allocated bandwidth, reduce min-guarantee M_{t_i} and reduce unsatisfied demands H_{t_i}. Reducing min-guarantee also reduces the tenant's bandwidth competing ability and hence increases H_{t_i}, resulting in utility decrease. Though increasing M_{t_i} strengthens a tenant's competing ability, it generates a much higher additional payment cost in our pricing model. Therefore, tenants are incentivized to limit their min-guarantee bandwidth to their exact needs. Bandwidth demand prediction [11], [25] can help tenants to estimate their demands. For a given demand $B^a_{t_i} = (B^c_{t_i}+B^u_{t_i})$, the payment cost is $\beta B^c_{t_i}+\gamma B^u_{t_i}$ $(\beta > \gamma)$; β is proportional to link congestion degree. Then, tenants are incentivized to use uncongested links instead of competing on congested links, to use less congested links and constrain link congestion. Consequently, with our network sharing policy, delay-tolerant applications may reduce unimportant demands or use less-important links to avoid bandwidth competition and congested links in order to pay less. The applications that compete for bandwidth are delay-sensitive applications, which however must pay high prices for their competed bandwidth. Then, the cloud achieves high overall performance for different delay-tolerant applications. These incentivized tenant behaviors benefit all tenants, increase network utilization and decrease unsatisfied demands, which increases the provider's utility.

In Section 3, we presented problems in the previous allocation policies: i) nodes are disincentivized to use uncongested links, and ii) nodes may cheat to gain more bandwidth allocation, both of which decrease network utilization. With our pricing model, tenants are incentivized to use uncongested links because they are cheaper than congested links; so problem i) is resolved. We then see if problem ii) is resolved. First, our definition of uncongested status is $D_l/C_l \leq 1$. If a link satisfies $D_l/C_l = 1$ (i.e., fully utilized), it is not congested, so it will not be considered in congestion proportionality. Thus, tenants on the links with $D_l/C_l = 1$ have no intention to reduce demands as it will not increase their allocation. If a link satisfies $D_l/C_l > 1$, it is congested and will be considered in congestion proportionality. Then, tenants are incentivized to reduce their demands to make the link satisfy $D_l/C_l = 1$ because of the cheaper unit price for uncongested links. This increases the utility of not only tenants but also the provider by reducing unsatisfied demands. Even though the tenant can gain more allocation, it still needs to pay for its gained additional bandwidth, which keeps proportionality.

Recall that when a bandwidth requesting VM v_i is notified about the possible congestion, it will try to reduce its demand to make the link uncongested. Then, it needs to make sure that its bandwidth demand reduction can increase utility; otherwise, it chooses not to reduce its demand. Assume that v_i needs to reduce x to make the link uncongested. Then, its utility equals $U^x_{v_i} = g_{t_i}(B^a_{t_i} - x) - [\alpha M_{v_i} + \gamma(B^a_{v_i} - x)] - F_{t_i}(x)$. If v_i chooses not to reduce demand, and it receives bandwidth $B^a_{v_i} - y$ after allocation, then its utility equals $U^y_{v_i} = g_{t_i}(B^a_{t_i} - y) - [\alpha M_{v_i} + \beta(B^a_{v_i} - y)] - F_{t_i}(y)$. Only when $U^x_{v_i} > U^y_{v_i}$, v_i can gain more utility by reducing x demand. Based on Equations (6) and (7), v_i can know its allocated bandwidth after allocation, denoted by $B^{ay}_{v_i}$. As a result, only when $g_{t_i}x + \gamma(B^a_{v_i} - x) + F_{v_i}(x) < g_{v_i}(B^a_{v_i} - B^{ay}_{v_i}) + \beta B^{ay}_{v_i} + F_{t_i}(B^a_{v_i} - B^{ay}_{v_i})$, v_i will reduce its demand by x to make the link uncongested.

4.4 Foreign Link Transmission

In this section, we propose a foreign link transmission policy to help reduce the number of congested links and unsatisfied demands in the multi-path topology in Figure 2. As the figure shows, each server connects to several switches, which allows the VMs in a server have multiple paths to the VMs in other servers. Each server has a local link that originally exists in the single-tree topology and multiple foreign links that are the additional links in the multi-tree topology.

In our pricing model, the VMs in each server are incentivized to choose the least congested links among the local links and foreign links to connect to other VMs, so that the probability of link congestion is significantly reduced. However, it is still possible that all of a VM's multiple paths have other VMs communicating on them and the least congested link cannot satisfy the VM's demand. Then, the link congestion will degrade the performance of the VM.

We notice that though the least congested link cannot satisfy the VM's demand, the sum of the available bandwidths of the first a few least congested links may satisfy the VM's demand. Accordingly, we propose the foreign link transmission policy to better handle the congestion problem. In this policy, when the least congested link among the multiple links cannot satisfy the bandwidth request of a VM, if the VM can partition its traffic, it chooses multiple least congested links so that the sum of the available bandwidth of these links can satisfy this VM's demand. For instance, assume the link capacity is 100Mbps. VM A_1 requests for 80Mbps with minimum bandwidth guarantee 20Mbps. It has four foreign links $L_1 - L_4$ with available bandwidth 40Mbps, 50Mbps, 60Mbps, and 70Mbps, respectively. If A_1 chooses the least congested link, its demand is still unsatisfied. Then, A_1 chooses the first and the second least congested links to communicate on, which are L_1 and L_2 in this example. In this case, the available bandwidth for A_1 will be 60+70=130Mbps, which can satisfy A_1's demand. This foreign link transmission policy can greatly help reduce the unsatisfied demands and decrease the number of congested links.

4.5 Bandwidth Allocation Enhancement

Based on previous bandwidth allocation policy, we further propose a bandwidth allocation enhancement strategy to satisfy each VM's demand as more as possible, i.e., reduce unsatisfied demand. For simplicity, we first take two VMs as an example and then extend it to multiple VMs. Assume two VMs v and w communicate on one link L with capacity C and they are competing for bandwidth allocation. v is a VM of tenant t_i, while w is a VM of tenant t_j. Then, the bandwidth demand on link L equals $D_l = B^a_{t_i,v} + B^a_{t_j,w}$ with minimum guarantee M_v and M_w, respectively.

According to our allocation policies based on Equations (6) and (7), the maximum bandwidth that can be allocated to VMs v and w are $V = C_l \frac{M_v}{M_v+M_u}$ and $W = C_l \frac{M_w}{M_v+M_w}$, respectively. Sometimes the link capacity is not fully utilized even though the sum of the two VMs' demands is higher than the capacity. To show this, we discuss the different bandwidth allocation cases below.

(1) If $B^a_{t_i,v} + B^a_{t_j,w} \leq C_l$, then the bandwidth allocation of VMs v and w will be just $B^a_{t_i,v}$ and $B^a_{t_j,w}$, respectively. In this case, there is no unsatisfied demand.

(2) If $B^a_{t_i,v} + B^a_{t_j,w} > C_l$, $B^a_{t_i,v} > V$ and $B^a_{t_j,w} > W$. In this case, since the link is congested, VMs v and w should be allocated with V and W, respectively. However, the demands are not satisfied for both v and w. Then, they may try to reduce their unimportant demands or choose more least congested links for partitioned traffic. In this case, the bandwidth allocation is fixed for the two VMs (i.e., V and W) and they receive bandwidths lower than their original demands.

(3) If $B^a_{t_i,v} + B^a_{t_j,w} > C_l$, $B^a_{t_i,v} < V$ and $B^a_{t_j,w} > W$. In this case, since the link is congested, according to our allocation policy, VMs v and w should be allocated with bandwidths V and W, respectively. However, the demand of VM v is $B^a_{t_i,v}$, which is smaller than its allocated V bandwidth. As VM v only needs to consume $B^a_{t_i,v}$, then there will be a waste bandwidth allocation $V - B^a_{t_i,v}$. To meet the requirement of high utilization, VM v should be allocated with bandwidth $B^a_{t_i,v}$ rather than V. As VM w demands more than its allocated bandwidth (i.e., $B^a_{t_j,w} > W$), VM w should be allocated with the remaining bandwidth $C_l - B^a_{t_i,v}$, which is greater than W, so its unsatisfied demanded bandwidth is reduced.

(4) If $B^a_{t_i,v} + B^a_{t_j,w} > C_l$, $B^a_{t_i,v} > V$ and $B^a_{t_j,w} < W$. It is a similar case as the above situation, so we do not repeat the discussion here.

In case (1) and case (2), the VMs either receive their demanded bandwidths or receive the possible maximum bandwidths lower than their demands, so that it is not necessary or possible for each of them to receive more bandwidth. However, in both case (3) and case (4), the link is already congested based on demands, but the total bandwidth usage (i.e., $B^a_{t_i,v} + W$ for case (3) and $V + B^a_{t_j,w}$ for case (4)) of two VMs is smaller than the link capacity. Hence, it is necessary to improve the bandwidth allocation policy in these cases to improve the resource utilization. In this bandwidth allocation enhancement strategy, in case (3), i.e., when VM v's demand $B^a_{t_i,v}$ is smaller than its allocated bandwidth V, the extra allocation $V - B^a_{t_i,v}$ will be taken out from VM v's allocation and allocated to the VMs with unsatisfied demands, i.e., VM w in this example. Finally, VM v receives its demanded $B^a_{t_i,v}$ bandwidth instead of V, while VM w is allocated with $C_l - B^a_{t_i,v}$ instead of W so that its unsatisfied demanded bandwidth is reduced. The same applies to case (4). This strategy increases link resource utilization and reduces unsatisfied demanded bandwidth of VMs.

For example, VM v and VM w are communicating on link L with capacity 1000Mbps. Assume the minimum guarantees of v and w are 200Mbps and 300Mbps, respectively, and their demands are 300Mbps and 800Mbps, respectively. Therefore, the link will be congested since the total requested bandwidth of VMs v and w is 300+800=1100>1000Mbps. Based on the bandwidth allocation enhancement policy, first, (V, W) are calculated, which are (400, 600)Mbps. Since $B^a_{t_i,v}$ = 300Mbps<V=400Mbps and $B^a_{t_j,w}$=800Mbps>W=600Mbps, this example is the case (3) stated above. Then, v's extra bandwidth $V - B^a_{t_i,v}$=100Mbps is taken out from v's allocation, and is allocated to VM w. Finally, the bandwidth allocation of VMs v and w are adjusted to 300Mbps and 700Mbps, respectively.

Note that in a realistic cloud network, generally there are a set of VMs $v_1, v_2, ..., v_n, w_1, w_2, ..., w_{n'}$ rather than only two VMs competing on one link. Assume that based on the original bandwidth allocation policy, $(V_1, V_2, ..., V_n)$ are the allocated bandwidths to VMs that demand lower bandwidths than the allocated bandwidths, and $(W_1, W_2, ..., W_{n'})$ are the allocated bandwidths to VMs that demand higher bandwidths than the allocated bandwidths. Suppose $\Delta V_k = V_k - B^a_{t_i,v_k} > 0$ and $\Delta W_k = W_k - B^a_{t_j,w_k} < 0$. Then, we can have two arrays: $\Delta\mathcal{V} =$

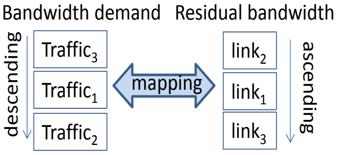


Fig. 5: Link mapping.

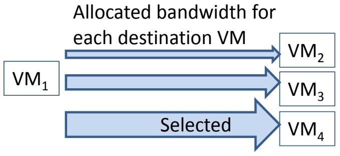


Fig. 6: Destination VM selection.

$(\Delta V_1, \Delta V_2, ..., \Delta V_n)$ and $\Delta\mathcal{W} = (\Delta W_1, \Delta W_2, ..., \Delta W_{n'})$. This bandwidth allocation enhancement policy distributes the extra bandwidth to the VMs with unsatisfied demands based on their weights calculated based on Equations (6) and (7). That is, a VM with a higher weight has a higher priority to receive the extra bandwidth. Specifically, this policy orders $\Delta\mathcal{V}$ in a descending order and order $\Delta\mathcal{W}$ in a descending order of the VMs' weights: $\Delta\mathcal{V} = (\Delta V^1, \Delta V^2, ..., \Delta V^n)$ and $\Delta\mathcal{W} = (\Delta W^1, \Delta W^2, ..., \Delta W^{n'})$. This strategy fetches the top element from $\Delta\mathcal{V}$ and allocates this amount of bandwidth to the VM of ΔW^1. Then, the two arrays are updated. This process repeats until $\Delta\mathcal{V}$ becomes empty. This bandwidth allocation enhancement policy greatly helps increase the link resource usage and reduce unsatisfied demands.

4.6 Traffic Flow Arrangement Policy

We use $\mathcal{PM} = \{p_1, p_2, \cdots\}$ to denote the set of all PMs in the cloud. Suppose tenant t_i has N_{t_i} VMs. The communication between these VMs can be represented by a matrix $\mathcal{M}_{N_{t_i}\times N_{t_i}}$. The value of each matrix element $U_{v_iv_j}$ means the utility of traffic traverse from VM v_i to VM v_j. The objective of a tenant is to maximize its utility U_{t_i}. To achieve this objective, the tenant aims to maximize $\sum\limits_{1\leq i\leq N_{t_i}} \sum\limits_{1\leq j\leq N_{t_i}} U_{v_iv_j}$. $U_{v_iv_j}$ equals:

$$g_{t_i}(B^c_{v_i,v_j}+B^u_{v_i,v_j})-(\alpha M_{v_i,v_j}+\beta B^c_{v_i,v_j}+\gamma B^u_{v_i,v_j})-F_{t_i}(H_{v_i,v_j}),$$

where the notations have the same meanings as before except they are for the VM-pair of v_i and v_j. We can formalize this objective as follows.

$$\max \sum_{j=1}^{N_{t_i}} \sum_{i=1}^{N_{t_i}} U_{v_iv_j} \tag{10}$$

$$\text{s.t. } B^a_{v_i,v_j} \leq Q_{v_i,v_j},\ \forall i,j \in \{1,...N_{t_i}\}, \tag{11}$$

$$B^a_{v_i,v_j} \geq L_{v_i,v_j},\ \forall i,j \in \{1,...N_{t_i}\}, \tag{12}$$

$$\sum_{v_i\in V_{p_m}} B^a_{v_i} \leq C_{p_m},\ \forall p_m \in \mathcal{PM}. \tag{13}$$

where V_{p_m} is the set of all VMs in PM p_m, $B^a_{v_i} = \sum\limits_{j=1}^{N_{t_i}} B^a_{v_i,v_j}$ is the total bandwidth of VM v_i. Q_{v_i,v_j} is the upper bound for bandwidth allocation from v_i to v_j, which is denoted by

$$Q_{v_i,v_j} = \min\{D_{v_i,v_j}, C_{p_m}, C_{p_n}\},\ s.t.\ v_i \in p_m, v_j \in p_n,$$

where C_{p_m} denotes the bandwidth capacity of PM p_m.

To achieve this objective, we let each VM distributively determine the links for its traffic flows. Each row in the matrix means v_i sends data to each v_j $(1 \leq j \leq N_{t_i})$. For a particular v_j, v_i can have multiple paths to send data to v_j [21]–[24]. We classify the flows that v_i attempts to send out into two types: *destined flow* and *non-destined flow*. A *destined flow* must traverse to a specified VM, while a *non-destined flow* can change its destination. For example, the data that is needed by a task executed in VM v_i is destined flow to v_i. The data of a computing task (e.g., WordCount) that can be assigned to any VM that has enough capacity to handle the task is non-destined flow.

TABLE 1: Denotation of each policy.

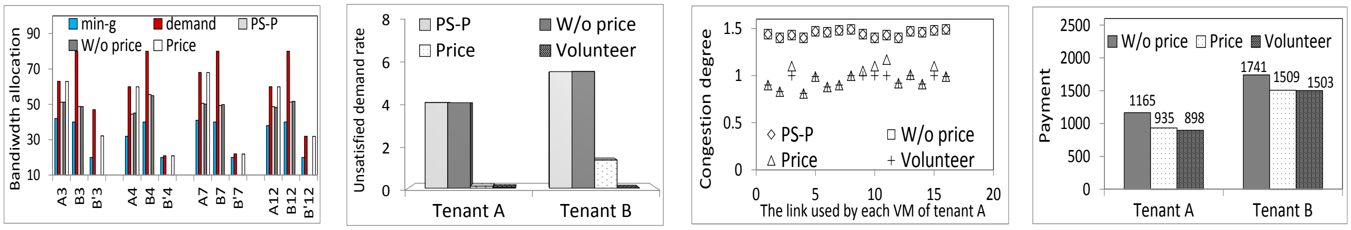
PS-P [3]	Minimum allocation
W/o price	Our allocation policy (i.e., PS-P+proportional policy)
Price	Our allocation policy+ pricing model
Volunteer	*Price*+ volunteering
Foreign	*Price*+ foreign link policy
Volunteer+Foreign	*Volunteer*+ foreign link policy
Enhanced	*Price* + bandwidth allocation enhancement

Recall that communicating along congested links are more expensive than communicating along uncongested links, while higher congested links are more expensive than less congested links. Therefore, v_i tries to choose the cheapest link (i.e., least congested) to traverse its destined flows. Always choosing the least congested link for each flow may not maximize $\sum_{1\leq j\leq N_{t_i}} U_{v_iv_j}$ globally because the residual bandwidth in the least congested link may be fragmented, which otherwise can support a high-demand flow. Failing to find a link to support a high-demand flow leads to competition. To handle this problem, we propose a link mapping algorithm as shown in Figure 5. v_i orders all destined flows based on bandwidth demand in descending order and orders the links based on residual bandwidth in ascending order. For each flow, v_i checks the link list in sequence until it finds one that has residual bandwidth no less than the flow's demand, and assigns this flow to this link. If a flow fails to find such a link, it is assigned to the last link with the maximum residual bandwidth, which can minimize the congestion degree. After each assignment, the two lists are updated. Using this way, the flows are assigned to links that have sufficient bandwidth to support the flow first or that lead to the least unsatisfied demand, thus increasing the utility of both the tenant and the provider. In the latter case, because of competition, the bandwidth allocation should be conducted.

If a flow is non-destined flow, v_i can assign it to any v_k $(1 \leq k \leq N_{t_i})$ that has enough capacity (i.e., CPU and storage) to handle the task of the flow. We then introduce the destination VM selection policy to help v_i gain more bandwidth. As shown in Figure 6, v_i can choose a VM that leads to the highest allocation based on Equations (6) and (7): $M_{v_i,v_j} + R\frac{W_{v_i,v_j}}{\sum_{v_m,v_n} W_{v_m,v_n}}$, where R is the residual bandwidth and v_m and v_n are the VM-pairs that are using the link's bandwidth. Consequently, each VM selects links and destinations for its flows to use uncongested links and constrain the congestion degree, which increases network utilization and reduce unfilled demands. For each flow transmission, the policy in Section 4.2 is used to prevent the occurrence of congestion, and allocate bandwidth based on min-guarantee and proportionality in congested links. The link mapping and destination VM selection policies help better arrange a VM's multiple flows to increase the utilities of both the tenants and the provider.

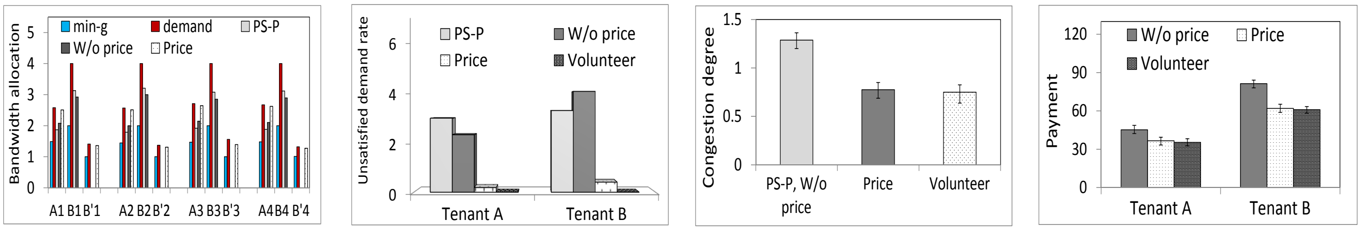
5 Performance Evaluation

We use simulation and trace-driven experiments to evaluate the performance of our proposed policies compared with previous bandwidth allocation policy. Specifically, we use PS-P [3] as the baseline. In order to see the contributions of our different policies, we tested different methods shown



(a) Bandwidth allocation. (b) Unsatisfied demand rate. (c) Congestion degree of links. (d) Payment of each tenant.

Fig. 7: Experimental results in simulation for different bandwidth allocation policies.



(a) Bandwidth allocation. (b) Unsatisfied demand rate. (c) Congestion degree of links. (d) Payment of each tenant.

Fig. 8: Trace-driven experimental results for different bandwidth allocation policies.

in Table 1. Recall that our proposed policies aim to satisfy each tenant's bandwidth demands as much as possible and increase network utilization by encouraging tenants to use idle links and avoid congested links. To show the benefits of our proposed polices, we tested the following metrics.

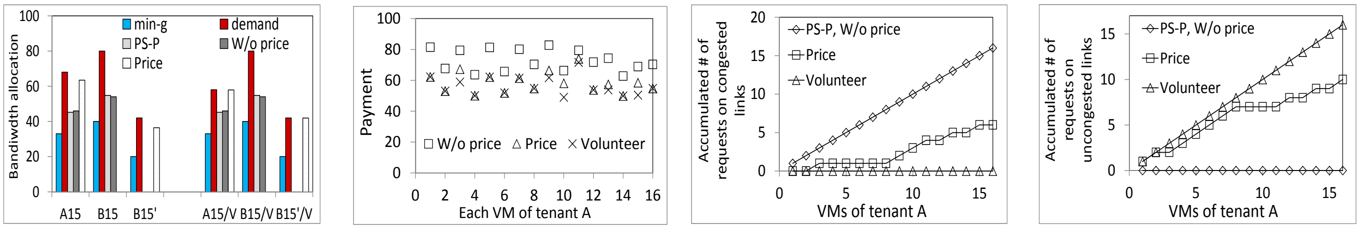
- *Bandwidth allocation.* It is the allocated bandwidth of each VM or a tenant. This metric shows whether each VM or tenant receives its min-guarantee bandwidth and demanded bandwidth, and how much demand is received.
- *Unsatisfied demand rate.* The *unsatisfied demand rate* of VM v_i is defined as $\frac{D_{v_i}-B^a_{v_i}}{D_{v_i}}$, and the *unsatisfied demand rate* for a tenant is defined as the sum of unsatisfied demand rate of each VM of the tenant: $\sum_{v_i\in V_t}\frac{D_{v_i}-B^a_{v_i}}{D_{v_i}}$. This metric shows the percentage of demanded bandwidth that is not satisfied. It also indicates the network utilization; a higher metric value means lower network utilization and vice versa.
- *Congestion degree.* It is calculated by $\frac{D_l}{C_l}$, where D_l denotes the total bandwidth demand on a link and C_l denotes the bandwidth capacity of the link. A higher metric value beyond 1 means the link is more congested, and a lower metric value below 1 means the link is more underutilized. Both cases mean low network utilization in the system given the same total demand amount. A metric value lower and close to 1 means that the link bandwidth is almost fully utilized and all demands on the link are satisfied.
- *Link utilization*. It is the percent of bandwidth being used by VMs communicating on the link. This metric measures whether the link capacity is fully utilized.

5.1 Performance of Our Pricing and Network Sharing Policies

As in [3], we use a tree topology as shown in Figure 2 [22] in our experiments. There are 16 servers and 2 tenants A and B in this scenario. Each tenant has one VM in each of the servers. Each server has a local link and three foreign links connecting to other VMs not in the same server. We assume tenant A's VMs communicate with its other VMs using a one-to-one communication pattern (i.e. $A_i \leftrightarrow A_{i+8}$, where $i = 1, 2, ...8$), while tenant B's VMs communicate with all of its other VMs (i.e $B_i \leftrightarrow B_j$, where $i \neq j$). Tenant B has two sets of VMs: B_i and B'_i $(1 \leq i \leq 16)$. Each B_i has 40Mbps minimum-guarantee and has already been allocated with 80Mbps bandwidth on each local link. Each B'_i has minimum bandwidth of 20Mbps and has been allocated with bandwidth randomly chosen from (0,100)Mbps on A_i's selected foreign link. Each of tenant A's VMs makes requests of bandwidth randomly selected from [60,70)Mbps and their minimum-guarantees are randomly selected from [30,40)Mbps. We set $\alpha = 1$ and $\gamma = 0.3$.

For the trace-driven experiments, we deployed Hadoop on a cluster running WordCount and then collected the transmitted and received bytes of each VM every second for 100 seconds. We use this trace in the experiments with the same settings as the above. In the experiment, each VM made requests for bandwidths based on the trace. We measured the metric each second for 100 seconds and present the median, the 95^{th} and 5^{th} percentiles of the metric results. In the figures, *min-g* represents the min-guarantee amount and *Demand* is the total bandwidth demand.

Figures 7(a) and 8(a) show the min-guarantees and demands of VMs, and their allocated bandwidths in different policies in simulation and trace-driven experiments, respectively. We show the results of 4 pairs of VMs rather than all 16 pairs to make the figure easy to read. Unlike *W/o price*, *Price* encourages VMs to choose the least congested links. Thus, in *W/o price*, VM A_x shares the link with B_x, while in *Price*, A_x shares the link with B'_x. Therefore, in the figures, B_x does not have the *Price* result and B'_x does not have the *W/o price* and *PS-P* results. We see that in *PS-P* and *W/o price*, tenant A's VMs with larger min-guarantees receive more bandwidth and vice versa. This is because the min-guarantees of B's VMs on the local links are fixed, and then tenant A's VM with a higher min-guarantee has a higher weight based on Equation (7), so it receives more bandwidth and hence tenant B's VM receives less bandwidth. We see that in *PS-P*, the VMs of tenant A and tenant B always cannot receive their demanded bandwidth. In *Price*, tenant A's VMs avoid using congested local links and are incentivized to use the least congested foreign links. Therefore, the VMs of A and B can gain their demanded bandwidth most of the time. In *Price*, we see that only B'_3 in simulation receives bandwidth less than its demand,



(a) VM A_{15}'s bandwidth allocation. (b) Payment of tenant A's VMs. (c) Accumulated number of requests on congested links. (d) Accumulated number of requests on uncongested links.

Fig. 9: Experimental results in simulation for different bandwidth allocation policies.

and in the trace-driven experiments, A_2 and A_4 receive their demanded bandwidth, while A_1 and A_3 are allocated with more bandwidth than in *PS-P* and *W/o price* but not all of their demands. This is because the least congested foreign link also becomes congested. Then, our allocation policy is employed to allocate bandwidth, which ensures min-guarantee first and then allocates the bandwidth based on the proportionality. These experimental results show the advantage of our pricing model to incentivize tenants to avoid bandwidth competition and fully utilize bandwidth resources, while ensuring min-guarantees.

In the experiments, the VMs with unsatisfied demands include A_3, A_9, A_{10}, A_{11}, A_{13} and A_{15}. In *Volunteer*, they volunteer to reduce their demands to make links uncongested. We use A_{15} as an example to show the details. Figure 9(a) shows the results of A_{15}, B_{15} and B'_{15} with and without the volunteer demand reduction employment in simulation. It shows that if A_{15} does not reduce its unimportant demand to make the link uncongested, both A_{15} and B_{15} cannot receive their demanded bandwidths. Otherwise, both receive their demanded bandwidth. This result confirms the effectiveness of our pricing model in incentivizing tenants to reduce their unimportant demands to make links uncongested.

Figures 7(b) and 8(b) show the unsatisfied demand rate in each method in simulation and trace-driven experiments, respectively. They indicate that without our pricing model, *PS-P* and *W/o price* only achieve different fairness in allocation but cannot prevent bandwidth competition. With our pricing model, *Price* reduces the unsatisfied demand for both tenants A and B because they are incentivized to select the least congested links in order to reduce payment. We see that *Volunteer* further reduces the unsatisfied demand rate for both tenants A and B. Tenant A volunteers to reduce its unimportant demand, which reduces the unit price for bandwidth consumption and SLO violations.

Figure 7(c) shows the congestion degree of each link used by each VM of tenant A in simulation. Figure 8(c) shows the median, 5^{th} and 95^{th} percentiles of the congestion degree of links used by tenant A in the trace-driven experiments. We see that without our pricing model (*PS-P* and *W/o price*), all links are congested. With our pricing model (*Price* and *Volunteer*), the congestion degree stays around 1. Since the unit price for uncongested links is lower than that of congested links, tenant A is incentivized to use uncongested links, leading to low link congestion degrees. *Volunteer* further reduces the congestion degree of the link by encouraging tenants to reduce unimportant demands to reduce unit price. For example, it reduces the link used by A_{15} from 1.1 in *Price* to 1. The results indicate the effectiveness of *Price* in maintaining the uncongested status.

We also show the payment in each policy in order to show the incentives provided by our policies. That is, tenants choose uncongested links or use our proposed policies because they can pay less, which increase network utilization. Figures 7(d) and 8(d) show the total payment of tenant A and tenant B (including VMs B_i and B'_i) in simulation and trace-driven experiments, respectively. For *W/o price*, even though it does not have the pricing model, we measured its payment using Equation (5) in order to show the incentives for the tenants in other policies to have less payment. The figure indicates that with our pricing model, if tenants use the less congested links, they pay less. We also see that *Volunteer* produces slightly less payment than *Price* for both tenants because some VMs reduce their unimportant demands. Figure 9(b) shows the payment of each VM of tenant A in different allocation policies. The figure shows that some VMs pay less in *Volunteer* than in *Price*. Also, VMs pay much less in *Price* than in *W/o price*. The results imply that VMs may volunteer to reduce their unimportant demands and they prefer to choose less congested links in order to reduce payment. These actions in turn reduce the link congestion, which reduces resource competition and benefits both the cloud provider and tenants.

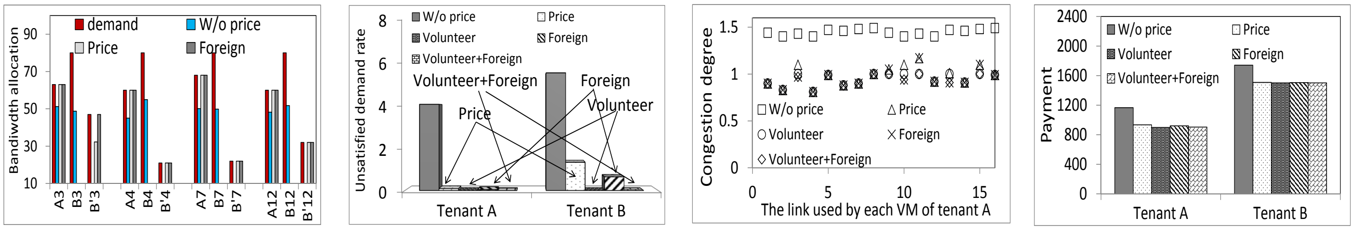
We then show the congestion status and payment of each VM in tenant A in the simulation in Figures 9(c) and 9(d). We see that the accumulated number of requests on congested links follows *PS-P=W/o price>Price>Volunteer*, and accordingly the accumulated number of requests on uncongested links follows *PS-P=W/o price<Price<Volunteer*. The results confirm that our allocation policies can help VMs avoid communicating on congested links and may further encourage tenants to reduce their unimportant demands to avoid congestion. We found that the trace-driven experimental results match the simulation results. We do not show the figures in this paper due to space limit.

In summary, our pricing model benefits both the provider and tenants. The tenants tend to pay less by using uncongested links and reducing their unimportant demands on congested links, which enhances the performance of tenant applications, increases the network utilization and reduces the SLO violations of the providers.

5.2 Performance of Foreign Link Transmission Policy

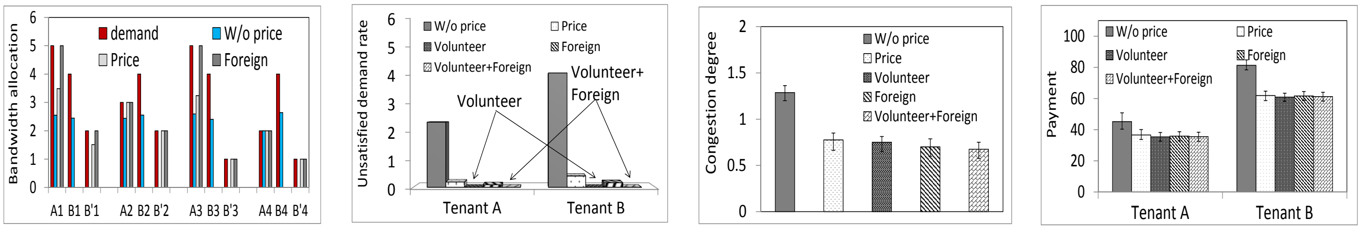
We have conducted the simulation of this policy with the aforementioned tree structure under the same experiment setting. Recall that VMs with divisible traffic can use the foreign link transmission policy, while the VMs with indivisible traffic cannot. We randomly let VMs with unsatisfied demands choose whether or not to use this policy. In the exper-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2015.2497701, IEEE Transactions on Parallel and Distributed Systems



(a) Bandwidth allocation. (b) Unsatisfied demand rate. (c) Congestion degree of links. (d) Payment of each tenant.

Fig. 10: Experimental results in simulation for the foreign link transmission policy.



(a) Bandwidth allocation. (b) Unsatisfied demand rate. (c) Congestion degree of links. (d) Payment of each tenant.

Fig. 11: Trace-driven experimental results for the foreign link transmission policy.

iments, the VMs with unsatisfied demands include A_3, A_9, A_{10}, A_{11}, A_{13} and A_{15}. Among these VMs, A_3, A_{10}, and A_{13} chose to use this policy, while the other three VMs did not.

Figure 10(a) and Figure 11(a) show the demand of four selected VMs and bandwidth allocation results in simulation and trace-driven experiments, respectively, with different allocation policies including *Foreign* and *Volunteer+Foreign*. We do not show the results of *min-g* and *PS-P* here since they are already displayed and compared in Figure 7(a) and Figure 8(a). In the figures, *Foreign(V)* means both methods of *Foreign* and *Volunteer+Foreign*. In Figure 10(a), in *Price*, the demand of VM B'_3 cannot be satisfied even though it communicates on the least congested link. This is because when A_3 selects the least congested link to communicate on, the link bandwidth is allocated between A_3 and B'_3, and then B'_3 receives less bandwidth than its demand. In *Foreign*, A_3 uses the foreign link transmission policy to transmit its traffic through two least congested links. Then, A_3 and B'_3 do not need to compete the link bandwidth, and B'_3's demand can be satisfied. As indicated previously, in *Volunteer*, the demands of both A_3 and B'_3 can be satisfied, since A_3 reduces its demand to make the link uncongested. Thus, *Foreign* can have more bandwidth allocation for the VMs, because it can satisfy their original demands, while *Volunteer* can only satisfy their reduced demands. The original demands of A_4, A_7 and A_{12} can be satisfied, so they do not need to use the foreign link transmission policy. A_3 does not need to volunteer to reduce its demand in *Foreign* as its demand is satisfied after employing this policy. As a result, there is no difference between the results in *Volunteer+Foreign* and in *Foreign* on the four VMs in the figure. In Figure 11(a), similarly, *Foreign* outperforms *Price* since *Foreign* can even satisfy the original demand of B'_1.

Figure 11(a) also indicates that the VMs in *Foreign* receive more bandwidth allocation than in *Price*. We see that in *Price*, the demands of both A_1 and A_3 cannot be satisfied even though they choose the least congested link. When A_1 and A_3 use the foreign link transmission policy by partitioning their demands to communicate on two least congested links, their demands are satisfied. Therefore, foreign link transmission policy further reduces unsatisfied demand rate.

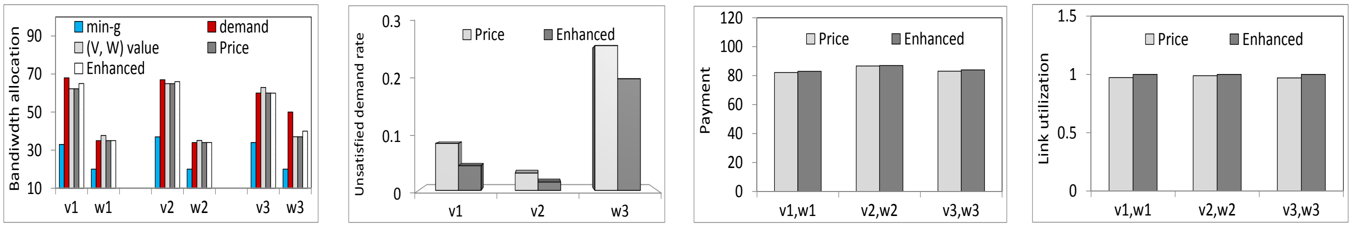
Figure 10(b) and Figure 11(b) show the unsatisfied demand rate of tenants A and B with different allocation policies, in simulation and trace-driven experiments, respectively. Both figures show that *Foreign* has a lower rate than *Price* and *Volunteer+Foreign* has a lower rate than *Volunteer*. This is because the foreign link transmission policy in *Foreign* and *Volunteer+Foreign* allows VMs with unsatisfied demands (i.e., A_3, A_{10}, and A_{13}) to also choose second least congested links to communicate on. The unsatisfied demand rate of *Foreign* is greater than that of *Volunteer* because VMs in *Foreign* do not reduce their unimportant demands.

As a VM may use multiple foreign links, we measured the congestion degree of each link used by a VM and used the maximum congestion degree as the congestion degree of the VM. Figure 10(c) shows the congestion degree of each VM of tenant A in different allocation policies. Figure 11(c) shows the median, the 5^{th} and 95^{th} percentiles of the congestion degree of links used by tenant A in the 100 results. As shown in these two figures, the congestion degree of each VM in *Foreign* is lower than that in *Price*, and approximately the same as *Volunteer* and *Volunteer+Foreign*.

Figure 10(d) shows the payment of tenants A and B with different allocation policies in simulation. Figure 11(d) shows the median, 5^{th} and 95^{th} percentiles of the payment of tenants A and B in the 100 results. In both figures, we see that the payment of both tenant A and tenant B follows *W/o price>Price>Foreign≈Volunteer+Foreign≈Volunteer*. In *Foreign*, VMs with unsatisfied demands can choose foreign links, which makes congested links in *Price* uncongested. Since uncongested links have a lower price, *Foreign* results in lower payment than *Price*. In *Volunteer* and *Volunteer+Foreign*, tenants reduce their demands to make links uncongested, so they pay less due to reduced bandwidth consumption and price. The experimental results confirm the effectiveness of the foreign link transmission policy.

5.3 Performance of Bandwidth Allocation Enhancement

We still implemented this experiment in the multi-tree structure with the same experimental settings. We did not implement the volunteer strategy in this experiment. The exper-



(a) Bandwidth allocation.

(b) Unsatisfied demand rate.

(c) Payment of each VMs pair.

(d) Link utilization of each VM pair.

Fig. 12: Experimental results in simulation for the bandwidth allocation enhancement policy.

iment results show that three link bandwidth competitions are in the situation that can use the bandwidth allocation enhancement policy. For simplicity, we denote the competing VM pair by (v, w), where v is a VM of tenant A and w is a VM of tenant B. Recall that (V, W) denotes the allocated bandwidth using the original bandwidth allocation policy.

TABLE 2: Effect of bandwidth allocation enhancement in simulation.

(v, w)	VM pairs	min-g	demand	(V, W)	Price	Enhanced
(v_1, w_1)	(A_{15}, B'_{15})	$(34, 20)$	$(68, 35)$	$(62, 38)$	$(62, 35)$	$(65, 35)$
(v_2, w_2)	(A_{13}, B'_{13})	$(37, 20)$	$(67, 34)$	$(65, 35)$	$(65, 34)$	$(66, 34)$
(v_3, w_3)	(A_{10}, B'_{10})	$(30, 20)$	$(60, 50)$	$(63, 37)$	$(60, 37)$	$(60, 40)$

TABLE 3: Effect of bandwidth allocation enhancement in trace-driven experiments.

(v, w)	VM pairs	min-g	demand	(V, W)	Price	Enhanced
(v_1, w_1)	(A_2, B'_2)	$(1.42, 1)$	$(4, 2)$	$(2.9, 2.1)$	$(2.9, 2)$	$(3, 2)$
(v_2, w_2)	(A_5, B'_5)	$(3, 1)$	$(7, 1)$	$(3.8, 1.2)$	$(3.8, 1)$	$(4, 1)$
(v_3, w_3)	(A_9, B'_9)	$(1.6, 1)$	$(2, 4)$	$(3.1, 1.9)$	$(2, 1.9)$	$(2, 3)$
(v_4, w_4)	(A_{11}, B'_{11})	$(1.8, 1)$	$(6, 1)$	$(3.2, 1.8)$	$(3.2, 1)$	$(4, 1)$

TABLE 4: Evaluation of VM pairs that used the bandwidth allocation enhancement policy in trace-driven experiments.

	Unsatisfied demand rate			Payment		Link utilization	
	Price	Enhanced		Price	Enhanced	Price	Enhanced
v_1	0.27	0.25	(v_1, w_1)	4.89	4.92	0.99	1
v_2	0.81	0.43	(v_2, w_2)	8.28	8.5	0.95	1
w_3	0.52	0.25	(v_3, w_3)	4.56	5.1	0.78	1
v_4	0.50	0.33	(v_4, w_4)	5.75	6.3	0.64	1

Figure 12(a) shows the bandwidth allocation of the three VM pairs, (v_1, w_1), (v_2, w_2), and (v_3, w_3), that used the bandwidth allocation enhancement policy in different bandwidth allocation policies. Table 2 illustrates the corresponding values. Take (v_1, w_1) for instance, in *Price*, v_1 is allocated with V=62Mbps. w_1 is allocated with W=38Mbps but it only used its demanded 35Mbps bandwidth. Thus, 3Mbps is wasted even though v_1 receives bandwidth lower than its demand. Therefore, in this case, *Price* cannot fully utilize link capacity. In *Enhanced*, this extra 3Mbps bandwidth is allocated to v_1 to increase its allocated bandwidth from 62Mbps to 65Mbps. Thus, v_1's demand can be more satisfied. Similarly, we see that with this enhancement policy, v_2 with 67Mbps demand can receive more bandwidth (66Mbps) than its originally allocated bandwidth (65Mbps), and w_3 with 50Mbps demand can receive more bandwidth (40Mbps) than its originally allocated bandwidth (37Mbps). Table 3 illustrates the results in the trace-driven experiment. We see that with the enhancement policy, VMs v_1, v_2, w_3 and v_4 can receive more bandwidths than what they are originally allocated. For example, w_3 with 4Mbps demand can be allocated with 3Mbps, which is more than its originally allocated bandwidth 1.9Mbps. The results confirm the effectiveness of the enhancement policy in satisfying more demands and increasing network utilization.

Figure 12(b) shows the unsatisfied demand rate of VMs that receive extra bandwidths (i.e., v_1, v_2 and w_3) in *Price* and *Enhanced* in simulation. Table 4 displays the unsatisfied demand rate for the VMs that receive extra bandwidths in the trace-driven experiments. We see that the VMs have lower unsatisfied demand rates in *Enhanced* than in *Price*. The results confirm the high effectiveness of the enhancement policy. The total payment of each VM pair in simulation and trace-driven experiments is shown in Figure 12(c) and Table 4, respectively. The figures and the table both indicate that the VM pairs need to pay slightly more with this policy. This is simply because they have received more allocation with this policy than without this policy.

We measured link utilization for the links of the VM pairs that used the enhancement policy in simulation and trace-driven experiments, as shown in Figure 12(d) and Table 4. The results indicate that *Enhanced* produces higher link utilizations than *Price*, which confirms the effectiveness of the enhancement policy in increasing network utilization.

TABLE 5: Bandwidth allocation w/ and w/o the link mapping policy.

	Min-g	Demand	W/o mapping	W/ mapping
VM1	5	10	6.7	10
VM2	20	40	26.7	40
VM3	50	100	66.7	100

TABLE 6: Performance with and without the link mapping policy.

	Unsatisfied demand rate	Cong. degree	Payment	Total # of cong. links
W/o mapping	0.3, 0.3, 0.3	1, 1, 1.5	10, 41, 103	1
W/ mapping	0, 0, 0	1, 1, 1	8, 32, 80	0

5.4 Effectiveness of Traffic Flow Arrangement Policy

Consider that a VM has three available links (l_1, l_2 and l_3) with capacities equal to 10Mbps, 40Mbps, 100Mbps, respectively. The VM needs to send data to three other VMs (VM_1, VM_2 and VM_3) with demands of 10Mbps, 40Mbps and 100Mbps, respectively. We assume that without our link mapping policy, VM1 will be allocated in priority, then VM2 and VM3. Without the link mapping policy, the allocation is $VM_1 \rightarrow l_3$, $VM_2 \rightarrow l_3$, and $VM_3 \rightarrow l_3$ because l_3 always has the most available bandwidth. With this policy, the allocation is $VM_3 \rightarrow l_3$, $VM_2 \rightarrow l_2$, and $VM_1 \rightarrow l_1$.

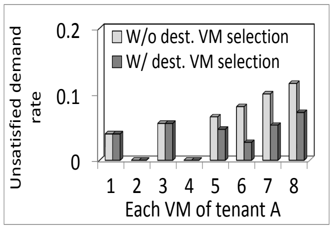


Fig. 13: Destination VM selection.

Table 5 and Table 6 show different metrics. In Table 5, we see that the bandwidth demand for all three destination VMs are satisfied with the policy, but are not satisfied without the policy. In Table 6, we see that this mapping policy reduces the unsatisfied demand rate, congestion degree and the payment for bandwidth usage, and the number of congested linked. The mapping policy globally considers the bandwidth demands and tries to satisfy each demand while avoids link congestion. More importantly, its payment reduction can incentivize tenants to carefully arrange their flows to different available links, which benefits both the provider and tenants. Figure 13 shows the unsatisfied bandwidth rate with and without our destination VM selection policy. We see that this policy is effective in reducing unsatisfied demands.

6 Conclusions

Network sharing in clouds is a critical issue in guaranteeing application performance. In this paper, we analyzed the behaviors of tenants in current pricing models and previously proposed bandwidth allocation policies in clouds. We found that these policies incentivize tenants to compete for bandwidth and even gain unfair allocation, which leads to low network utilization and degrades the benefits of both the cloud provider and other tenants. We then propose bandwidth sharing and pricing policies to transform the competitive environment to a win-win cooperative environment, where tenants strive to increase their utility, which also concurrently increases the utilities of the cloud provider and other tenants. Specifically, we propose a new bandwidth pricing model, a network bandwidth sharing policy, foreign link transmission policy, bandwidth allocation enhancement policy and flow arrangement policy. These policies incentivize tenants to use uncongested links and constrain congestion, and finally reduce unsatisfied demands and increase network utilization. The bandwidth allocation on congested links also meets the three desired requirements (min-guarantee, high utilization, and network proportionality) – an unsolved problem in previous research. Our experiments show the effectiveness of our proposed policies. In our future work, we will consider rewarding tenants for reducing demand to maintain the uncongested link states.

Acknowledgements

This research was supported in part by U.S. NSF grants NSF-1404981, IIS-1354123, CNS-1254006 and Microsoft Research Faculty Fellowship 8300751. An early version of this work was presented in the Proc. of INFOCOM 2014 [26].

References

[1] F. Xu, F. Liu, H. Jin, and A. V. Vasilakos. Managing performance overhead of virtual machines in cloud computing: A survey, state of the art, and future directions. *Proceedings of the IEEE*, 2014.

[2] A. Shieh, S. Kandula, A. Greenberg, C. Kim, and B. Saha. Sharing the data center network. In *Proc. of NSDI*, 2011.

[3] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, and I. Stoica. Faircloud: sharing the network in cloud computing. In *Proc. of SIGCOMM*, 2012.

[4] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron. Towards predictable datacenter networks. In *Proc. of the SIGCOMM*, 2011.

[5] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang. Secondnet: a data center network virtualization architecture with bandwidth guarantees. In *Proc. of CoNEXT*, 2010.

[6] J. Guo, F. Liu, D. Zeng, J. Lui, and H. Jin. A cooperative game based allocation for sharing data center networks. In *Proc. of INFOCOM*, 2013.

[7] L. Popa, P. Yalagandula, S. Banerjee, J. C. Mogul, Y. Turner, and J. Santos. Elasticswitch: practical work-conserving bandwidth guarantees for cloud computing. In *Proc. of SIGCOMM*, 2013.

[8] V. Jeyakumar, M. Alizadeh, D. Mazières, B. Prabhakar, Al. Greenberg, and C. Kim. Eyeq: Practical network performance isolation at the edge. In *Proc. of NSDI*, 2013.

[9] K. LaCurts, J. C. Mogul, H. Balakrishnan, and Y. Turner. Cicada: Introducing predictive guarantees for cloud networks. In *Proc. of HotCloud*, 2014.

[10] J. Lee, Y. Turner, M. Lee, L. Popa, S. Banerjee, J. Kang, and P. Sharma. Application-driven bandwidth guarantees in datacenters. In *Proc. of SIGCOMM*, pages 467–478, 2014.

[11] D. Niu, C. Feng, and B. Li. Pricing cloud bandwidth reservations under demand uncertainty. In *Proc. of SIGMETRICS*, 2012.

[12] Y. Feng, B. Li, and B. Li. Bargaining towards maximized resource utilization in video streaming datacenters. In *Proc. of INFOCOM*, 2012.

[13] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron. Better never than late: meeting deadlines in datacenter networks. In *Proc. of SIGCOMM*, 2011.

[14] J. Guo, F. Liu, J. Lui, and H. Jin. Fair network bandwidth allocation in iaas datacenters via a cooperative game approach. *TON*, 2015.

[15] J. Guo, F. Liu, X. Huang, J. Lui, M. Hu, Q. Gao, and H. Jin. On efficient bandwidth allocation for traffic variability in datacenters. In *Proc. of INFOCOM*, 2014.

[16] F. P. Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8(1):33–37, 1997.

[17] F. P. Kelly, A. K. Maulloo, and D. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, pages 237–252, 1998.

[18] J. K. Mackie-Mason and H. R. Varian. Pricing the Internet. Technical report, Oslo University, Department of Economics, 1993.

[19] M. J. Osborne and A. Rubinstein. *A course in game theory*. The MIT Press, July 1994.

[20] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *Proc. of SIGCOMM*, 2008.

[21] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley. Improving datacenter performance and robustness with multipath tcp. In *Proc. of SIGCOMM*, 2011.

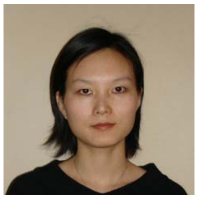
[22] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. Vl2: a scalable and flexible data center network. In *Proc. of SIGCOMM*, 2009.

[23] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu. Bcube: A high performance, server-centric network architecture for modular data centers. In *Proc. of SIGCOMM*, 2009.

[24] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul. Spain: Cots data-center ethernet for multipathing over arbitrary topologies. In *Proc. of NSDI*, 2010.

[25] D. Xie, N. Ding, Y. C. Hu, and R. R. Kompella. The only constant is change: incorporating time-varying network reservations in data centers. In *Proc. of SIGCOMM*, 2012.

[26] H. Shen and Z. Li. New bandwidth sharing and pricing policies to achieve a win-win situation for cloud provider and tenants. In *Proc. of INFOCOM*, 2014.



Haiying Shen received the BS degree in Computer Science and Engineering from Tongji University, China in 2000, and the MS and Ph.D. degrees in Computer Engineering from Wayne State University in 2004 and 2006, respectively. She is currently an Associate Professor in the Department of Electrical and Computer Engineering at Clemson University. Her research interests include distributed computer systems and computer networks, with an emphasis on P2P and content delivery networks, mobile computing, wireless sensor networks, and cloud computing. She is a Microsoft Faculty Fellow of 2010, a senior member of the IEEE and a member of the ACM.

Zhuozhao Li received the BS degree in Optical Engineering from Zhejiang University, China in 2010, and the MS degree in Electrical Engineering from University of Southern California in 2012. He is currently a Ph.D. candidate in Electrical and Computer Engineering at Clemson University. His research interests include data analysis, cloud computing and big data processing.