Modeling and Pricing Cloud Service Elasticity for Geographically Distributed Applications

Bassem Wanis, Nancy Samaan, Ahmed Karmouch
SITE, University of Ottawa, Ottawa, Canada
Email: bwanis@uottawa.ca, {nsamaan, karmouch}@site.uottawa.ca

***Abstract*—Cloud service providers (CSP) strive to effectively provision their cloud resources to ensure that their hosted distributed applications meet their performance guarantees. However, accurately provisioning the inter-data centers network resources remains a challenging problem due to the cloud hosted applications' workload fluctuation. In this paper, we propose a novel approach that enables a CSP to offer *Elasticity-as-a-Service* (EaaS) for inter-data centers communication in order to guarantee the performance of distributed cloud applications. The contributions of the proposed work are two fold; first, we develop an efficient approach that enables the CSP to estimate and reserve the pool of network resources needed to fulfill the demands imposed by the network workload fluctuations of applications subscribing to this service. The approach allows the CSP to offer communication EaaS at differentiated levels based on the degree of bandwidth-sensitivity of the distributed cloud applications. In order to capture the inter-data centers network activity of hosted applications, we model their workloads using Markovian modeling. The second contribution is a novel dynamic pricing mechanism for network EaaS offerings that can be employed by the CSP to maximize the expected long-term revenue, and to regulate network elastic demands. Performance evaluation results demonstrate the efficiency of our proposed approach, the higher accuracy of our prediction method, and the increase in the CSPs net profit.**

***Index Terms*—cloud elasticity; distributed cloud computing; network virtualization; inter-data centers communication; resource pricing.**

I. Introduction

The ability of cloud computing environments to offer scalable and cost-efficient computing and networking resources has contributed to the growth of large-scale geographically distributed applications. These applications are hosted on clouds and use dedicated virtual machines (VMs) residing on physical servers that are dispersed throughout multiple data centers [1], [2]. To facilitate the communication among these distributed VMs, CSPs own or lease a high-capacity backbone network[1] to carry the tenants traffic between data centers [2]–[4]. Network virtualization techniques [5], [6] are then employed for the network resource management as shown by Fig. 1. Unfortunately, the majority of current distributed cloud-based applications are long-lived services and characterized by a high degree of workload fluctuation. In turn, these applications require a continuous down- or up-scaling of the amount of allocated resources to accurately reflect the time-varying application's needs [7]–[9]. This fluctuation is largely attributed to the nature of the offered services and/or other external events that may result in incremental growth or sudden variation in popularity (e.g., releasing of a new movie, publishing of an article about a particular company on a highly visited news site, popularity participating in a conference).

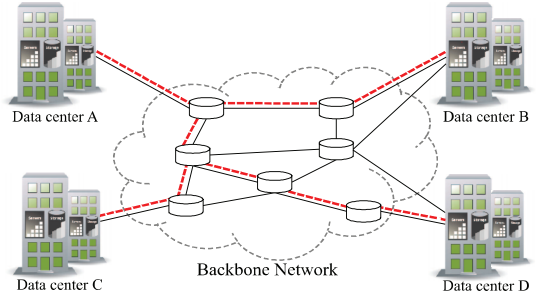

Fig. 1. An example of an inter-data centers backbone network

Cloud service elasticity refers to the ability of the CSP to dynamically allocate the resources according to the cloud application needs. Nonetheless, one of the limitations of the current cloud resource provisioning techniques is that they mainly focus on efficient and dynamic allocation of computational and network resources within data centers [10], [11], without paying much attention to inter-data centers network provisioning which highly affects the performance of these distributed applications. Moreover, VMs replication/migration and internal background processes (e.g., periodical data backup, data indexing, database distribution and synchronization, video encoding and caching, and MapReduce jobs) result in a huge amount of data being transferred among different data centers [8]. Therefore, the inter-data centers traffic load varies significantly, which stresses the need for inter-data centers network EaaS, a value-added service where the CSP allows the hosted applications to down- or up-scale its consumed resources in a continuous manner. Existing approaches are limited to either provisioning unbounded network resources, or statically provisioning the expected peak-traffic value. Both solutions lead to inefficient resource utilization resulting in performance degradation. So far, however, very few inter-data centers dynamic network provisioning techniques have been developed (e.g., [12]–[14]). Nonetheless, these techniques do not suggest a way that guarantees the availability of network resources in response to workload changes. Moreover, these techniques either assume that the time-varying probability distribution of cloud applications is well known, or ignore the repetitive behavior of long-lived cloud applications, resulting in inefficient provisioning approaches.

To overcome the aforementioned problems, we propose a

[1] Throughout this paper, whenever we mention network/communication resources, we refer to inter-data centers network/communication resources.

novel inter-data centers network resource management approach that dynamically re-sizes the inter-data centers bandwidth pool in order to ensure the availability of network resources. This enables the CSP to offer the promised network EaaS.

The CSP decides for all the distributed applications that have the same traffic workload patterns and similar degree of bandwidth-sensitivity, a pre-defined elasticity level that ensures their performance objectives. Based on the decided elasticity level and the expected traffic workload fluctuation, the CSP estimates the amount of network resources needed to be reserved for the next time interval[2]. After estimating the required EaaS pool size, the CSP either sets apart a portion of the network resources, if available, or it can outsource cloud resources across multiple CSPs [16]. In order to have an accurate prediction, we model the inter-data centers traffic workload using a Markov chain model, taking advantage of the temporal variability of workload fluctuations [17]. Also, this model can easily and accurately capture the repetitive workload behavior of long-term stationary applications (e.g., Netflix) in a straight forward manner and within a reasonable time [9], [17], [18]. The remainder of this paper is organized as follows. In Section II, related work and existing cloud elasticity techniques are briefly discussed. Sections III and V discuss our proposed network elasticity model and EaaS pricing mechanism, respectively. Simulation and performance evaluation results are discussed in Section VI. Finally, Section VII concludes this paper.

II. Related work

To date, the majority of the existing techniques in the literature that target elasticity for the infrastructure-as-a-service (IaaS) model, focus on computing, storage, and network resources within the same data center, with little attention to inter-data center network resources elasticity. Several commercial and academic elasticity approaches offer elasticity features that can be categorized as either rule-based reactive approaches, or predictive approaches. The former defines the set of actions that should be taken when one or multiple observed parameters (e.g., CPU usage, network traffic, disk access) exceed/fall-behind a pre-defined threshold. The latter adjusts the allocated resources based on the prediction of future violations and workload forecasts. *Auto-Scaling* offered by Amazon web services (AWS) [19] is based on a rule-based reactive approach that adds/releases cloud instances based on monitored parameters. Lim et al. [20] proposed a rule-based reactive approach where elasticity actions take place according to a target resource demand range, rather than a single target value, to avoid resource thrashing (i.e., constantly adding/releasing cloud instances). Dawoud et al. [21] proposed a dynamic resource provisioning approach that aims at minimizing the resources required to handle the future workload by following a linear prediction approach that decides the next allocation based on the the last allocation and consumption. Roy et al. [22] proposed a predictive approach for workload forecasting that aims at minimizing the resource provisioning costs while guaranteeing the application QoS. Gong et al. [9] proposed *PRESS*, a predictive elasticity scheme that is based on resource demand patterns prediction through Fast Fourier Transform (FFT), to minimize the unused resources while preventing service level objectives (SLO) violations. Sharma et al. [23] proposed *Kingfisher*, an elasticity provisioning approach that takes into consideration workload fluctuation while trying to minimize the virtual resources costs. *Kingfisher* computes the optimal resource provisioning based on the cost of migrating and/or replicating VM instances, also considering the transition time to allocate more resources. Rao et al. [7] proposed a self-tuning fuzzy control rule-based multi-objective resource allocation approach that guarantees the cloud applications QoS. Netflix is using *Scryer* [24], a predictive auto-scaling engine to deliver the best quality of experience (QoE) to Netflix customers. *Scryer* predicts upcoming workload based on two prediction algorithms, namely, an augmented linear regression-based and an FFT based algorithms. Li et al. [8] proposed a resource management approach that consider future demand fluctuations while provisioning the physical machines and links capacities.

Research efforts that are related to data centers network provisioning mainly focus on intra-data center network provisioning. Guo et al. [25] proposed *SecondNet*, a low time-complexity heuristics-based algorithm that allocates a virtual data center using clustered neighboring servers. Current demands are met by increasing or decreasing the bandwidth reservation along existing paths or through migrating virtual links to new paths. Benson et al. [26] proposed *CloudNaaS*, a framework that supports the deployment and management of cloud enterprise applications. The architecture supports both best-effort and bandwidth reservation based services among VMs. Nonetheless, the increase in the number of paths with bandwidth reservation may lead to congestion, starvation of other services in addition to an inefficient network utilization. Another bandwidth allocation technique, termed *Seawall* [10], performs end-to-end proportional sharing of the network resources according to an associated weight with each traffic source using congestion-controlled tunnels. Similarly, Lam et al. [11] proposed *Netshare*, a statistical multiplexing mechanism that proportionally allocates bandwidth for the tenants. Divakaran et al. [27] proposed a probabilistic-bandwidth model where bandwidth requirements are specified with some probabilities to overcome the weakness of deterministic models. Also, Niu et al. [12] proposed a bandwidth allocation model which enables the service providers (SPs) to simply specify a percentage of their demands to be allocated with a certain guarantee level (i.e., guaranteed portion).To date a small number of research efforts have addressed the problem of dynamic resource provisioning of inter-data centers connectivity. Of those efforts is the work of Ajay et al. [2] that proposed an architecture to enable the dynamic configuration of the inter-data centers network by offering a bandwidth-on-

[2] The duration of the time intervals is specified by the CSP and depends on the nature of the hosted distributed cloud applications [15]

demand service. Ghosh et al. [3] proposed a inter-data centers traffic management design by optimizing the sending rates and controlling the network routing, across multiple application traffic classes. Carella el al. [16] proposed an elasticity engine that is based on brokerage of cloud resources among multiple CSPs, in order to optimally allocate the cloud resources in federated cloud environments. In general, the aforementioned approaches do not consider the problem of the availability of inter-data centers network resources and how to estimate the pool size that needs to be reserved in order to offer the promised differentiated network EaaS. In addition, they do not consider the case of differentiated applications sensitivity and how to benefit from the demand repetitive patterns.

III. A Novel Elastic Service Offerings

One of the CSP's objectives is to effectively manage the underlying physical network by determining what portion of the network resources should be reserved to offer the EaaS, and the portion that may be sold to serve new cloud requests. To answer this question, the CSP needs first to accurately estimate the expected network elastic demands, then compares the trade-off between the long-term revenue earned from the EaaS, and the revenue from accepting new cloud requests. The CSP must also take into consideration the penalties that may be imposed due to the failure of meeting the elasticity service level due to the unavailability of sufficient network resources. Before describing our proposed approach, we will first propose an inter-data centers communication EaaS model.

A. A Novel EaaS model

The CSP considers each inter-data centers link as a separate network resource and estimates the bandwidth amount needed to be reserved according to the expected fluctuation of the virtual links mapped over it. To this end, the CSP's distributed cloud can be modeled as a weighted undirected graph $G(N, L)$, where N and L represent the sets of data centers and inter-data centers links, respectively as shown by Fig. 1. We use c_e to denote the capacity of $e \in L$, and x_e to denote the EaaS reserved bandwidth pool size of $e \in L$. Each distributed application v is hosted on a set of distributed VMs connected through the inter-data centers network and can be modeled as a weighted undirected graph $G^v(N^v, E^v)$, with the sets of virtual nodes N^v (i.e., VMs, virtual switches, and virtual routers), and virtual edges E^v hosted on an inter-data centers link e. The elastic bandwidth demand of each link $e^v \in E^v$ is denoted by d^v (i.e., d^v is the requested excess of bandwidth over the originally contracted values). Each distributed application specifies its service level objectives SLO which requires a specific level of network elasticity as will be shown next. According to the SLO, the CSP allocates the required resources needed to guarantee the application performance objectives. Beside the elasticity of the allocated distributed VMs resources (i.e., CPU, storage, and memory), the CSP considers the elasticity of the inter-data centers network to be able to deal with traffic workload fluctuations of each virtual link hosted on the network.

Our proposed EaaS model allows the CSP to offer differentiated elasticity services to its distributed cloud applications. The offered elasticity levels vary from perfect elasticity (i.e., 100% elasticity level means that the virtual link bandwidth appears to be unlimited) to partial elasticity (i.e., resources are guaranteed up to a certain level). The elasticity level is determined based on the nature of the distributed cloud application and the current stage of the application life-cycle. In general, applications that belong to the same application type and share the same time-varying behavior have the same network workload demands [28] (e.g., distributed video streaming cloud applications that serve the same area have the same traffic fluctuation and distribution parameters). According to that, all the distributed applications that have the same traffic workload patterns and similar degree of bandwidth-sensitivity are assumed to be categorized into the same class and are assigned the same elasticity level. Then the CSP calculates the required network resource pool size for this class according to its elasticity level. For simplicity, we will focus our analysis on a single EaaS class and leave the case of multiple classes for future work. The CSP estimates the bandwidth pool size $x_e(t)$ on each $e \in L$ for the n cloud applications that belong to the same class at a time. Each EaaS class is characterized by a probability $1-\alpha$ which represents the probability of elasticity service (i.e., α is the probability that one or more of the applications in that class will request additional resources and that the CSP will fail to satisfy the request). In other words, the value α represents the probability that the total bandwidth elasticity demands $\sum_{j=1}^{n} d^{v_j}(t)$ for the n cloud applications that belong to the same EaaS class, during a specific time interval t exceed the EaaS reserved pool $x_e(t)$ for that class at $e \in L$, i.e.,

$$Pr(\sum_{j=1}^{n} d^{v_j}(t) \geq x_e(t)) = \alpha \quad \forall e \in L \tag{1}$$

Clearly, α is inversely proportional to the contracted EaaS level. For instance, bandwidth-sensitive distributed cloud applications (e.g., video streaming) need a high EaaS level, while bandwidth-insensitive distributed cloud applications (e.g., offline laboratory experiments), which are able to tolerate bandwidth under-provisioning, may require a low elasticity level.

B. Application Demand Modeling

Predicting the varying inter-data centers communication demands of distributed cloud hosted applications is a necessary step in order to enable the CSP to accurately reserve the resources required to provide the required EaaS. The CSP can determine the expected bandwidth demands by monitoring the workload patterns over time of the distributed applications [9], [17]. Accordingly, a discrete-time Markov chain model is built based on historical network workload patterns to be able to estimate the short-term bandwidth demands. The state model is first derived by observing the application operation and stages, and then converted into a Markov chain representation in which each state of the model represents the time-varying probability distribution of the inter-data centers traffic for a given EaaS class. Long-term distributed cloud applications are stationary and characterized by repeated patterns for the demand fluctuation[3] which emphasize the suitability of using Markov modeling [30]. We consider a Markov chain model consisting of a finite set of states, $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$, each state $s_i \in \mathcal{S}$ is characterized by a mean $\mu(s_i)$ and a variance $\sigma^2(s_i)$ that are considered as the main parameters that describe the probability distribution of the traffic workload of a given EaaS class at this state [12], [14], [17]. At each period $t = 1, 2, \ldots$, the probability distribution parameters can be inferred from the current state $s_i \in \mathcal{S}$ that follows a Markov process characterized by a state transition matrix $\Pi = [\pi(s_i|s_j)]$, where $s_i, s_j \in \mathcal{S}$. The process starts at a specific state and moves gradually from one state $s(t-1)$ at time $t-1$ to another state $s(t)$ at time t, with a probability denoted by $\pi(s_j|s_i)$ as follows,

$$\pi(s_j|s_i) = Pr(s(t) = s_j|s(t-1) = s_i) \tag{2}$$

The conditional probability $\pi(s_j|s_i)$ follows a Markov process since the next state s_j is dependent only on the current observed state s_i, and is independent of the sequence of states that preceded it. Formally, it can be written as,

$$\begin{aligned} Pr(s(t) = s_j|s(t-1) = s_i, \ldots, s(1) = s_1) \\ = Pr(s(t) = s_j|s(t-1) = s_i) \end{aligned} \tag{3}$$

The state transition coefficients have the properties that $\pi(s_i|s_i) \geq 0$ and $\sum_{j=1}^{k} \pi(s_i|s_j) = 1$. The CSP estimates the mean and the variance of the probability distribution of the network traffic in discrete time intervals according to the state transition matrix Π given by the first-order Markov chain model inferred from historical monitoring. Based on these parameters, the CSP calculates the reserved bandwidth pool size that is needed to provide EaaS for the hosted applications at run-time.

IV. EaaS Reserved Resource Pool Calculation

At each time interval t and for a given EaaS class, the CSP derives the estimated state $s(t) = s_t$ of the Markov chain process for each link e, and determines the mean $\mu(s_t)$ and the variance $\sigma^2(s_t)$ of the bandwidth demands distribution. The bandwidth demand $d^{v_j}(t)$ for an application j, $j = 1, \ldots, n$, is considered as an independent and identically distributed (*i.i.d*) random variable, with mean $\mu = E(d^{v_j})$ and variance $\sigma^2 = Var(d^{v_j})$. According to the *Central Limit Theorem* [31], since the number of hosted cloud applications n is sufficiently large, the distribution of the arithmetic mean $\bar{d_n}(t) = (d^{v_1}(t) + d^{v_2}(t) + \cdots + d^{v_n}(t))/n$ of the bandwidth demand at state s_t can be approximated as a normal distribution with mean $\mu(s_t)$ and variance $\sigma^2(s_t)/n$, regardless of the actual unknown distribution of the bandwidth demand.

Hence, the distribution of the sum of the n random variables $\sum_{j=1}^{n} d^{v_j}(t)$ can also be approximated by a normal distribution with mean $n\mu(s_t)$ and variance $n\sigma^2(s_t)$. Formally, let $D(t) = \sum_{j=1}^{n} d^{v_j}(t)$, $\mu_D(t) = E[D(t)] = n\mu(s_t)$, and $\sigma^2_{D(t)} = Var(D(t)) = n\sigma^2(s_t)$. The CSP calculates the resource pool size $x_e(t)$ at each link e to be reserved at the required elasticity level α, so that the probability that the total bandwidth demands in that EaaS class may exceed the total reserved pool is less than or equal the predefined value α, i.e., $Pr(D(t) \geq x_e(t)) \leq \alpha$. According to the *Central Limit Theorem*, we have

$$\frac{D(t) - \mu_{D(t)}}{\sigma_{D(t)}} \sim N(0, 1). \tag{4}$$

So the estimated pool size at state s_t for each link e can be calculated to satisfy the constraint:

$$Pr(\underbrace{\frac{D(t) - \mu_{D(t)}}{\sigma_{D(t)}}}_{N(0,1)} \geq \underbrace{\frac{x_e(t) - \mu_{D(t)}}{\sigma_{D(t)}}}_{z_{1-\alpha}}) \leq \alpha \quad \forall e \in L \tag{5}$$

where $z_{1-\alpha}$ is the $1-\alpha$ percentile of the *Standard Normal* distribution. Thus, $x_e(t)$ can now be calculated as follows,

$$x_e(t) = n\mu(s_t) + z_{1-\alpha}\sqrt{n}\sigma(s_t) \tag{6}$$

It is worth noting, that this calculation is carried out per each state s_t for each EaaS class.

V. Proposed EaaS Pricing Model

In this section, we propose an elasticity dynamic pricing model that aims at maximizing the CSP expected long-term revenue. In contrast to the traditional static pricing models which are not tailored to the elastic behavior of cloud traffic workload, the main goal of this model is to provide the optimal price vector obtained from offering EaaS that maximizes the expected long-term revenue. Let the price vector $\mathbf{P}_e = (p_e(s_1), \ldots, p_e(s_k))$ provides the prices of the EaaS link e for each state $s_i \in \mathcal{S}$. The CSP calculates the price vector that maximizes its revenue during an operating time cycle $[0, T]$. This price vector is updated periodically to reflect any variations that have been occurred to the EaaS demands patterns. Cloud applications are price-sensitive, in other words, when the prices increase, cloud applications owners seek to reduce the workload to the cloud which directly affects the workload traffic distribution. On the contrary, when the prices decrease, cloud applications owners will be encouraged to move more of their workload to the cloud. At each period $t \in [0, T]$, the bandwidth pool size $x_e(t)$ is reserved to fulfill the total bandwidth fluctuation per EaaS class which is calculated based on the current state $s(t) = s_t$ as shown in the previous section. A scaling-up in the link total demand happens at time t, if $dx_e(t) > 0$, while $dx_e(t) < 0$ indicates a scaling-down ($dx_e(t) = 0$ means no fluctuation). At any

[3] The CAIDA Equinix inter-data centers traces [29] show the repeated behavior of inter-data centers traffic, as will be seen in section VI.

time t the total bandwidth demand should not exceed the link capacity c_e by satisfying this condition:

$$\int_0^t dx_e(t) \in [-x_e(0), c_e - x_e(0)], \qquad \forall t \in [0, T], \quad (7)$$

where, $x_e(0)$ represents the reserved bandwidth pool size at the beginning of the interval $[0, t]$. Moreover, the prices should be bound by a maximum and a minimum values, p^{max} and p^{min} respectively, $p_e(s_t) \in [p_e^{min}, p_e^{max}]$ $\forall t \in [0, T]$.

Since, the traffic workload fluctuates over the period T, the CSP aims at finding the price vector that maximizes its revenue. The long-term excepted revenue from a given EaaS class at link e can be calculated by,

$$\mathcal{R}_e(T) = \int_0^T p_e(s_t) x_e(t) dt \quad (8)$$

Now, the problem is to find the optimal price vector $\mathbf{P}_e^*$ that maximizes the CSP revenue denoted by $\mathcal{R}_e^*(T)$, and at the same time, reflects the link utilization level to prevent the under-utilization/congestion by decreasing/increasing the price vector of each link, respectively. The maximum CSP revenue $\mathcal{R}_e^*(T)$ can be considered as the *least upper bound* (lub) of $\mathcal{R}_e(T)$ that satisfies the prices boundary constraint. To search for the optimal price vector, we continuously apply dynamic programming algorithm backwards in time [32]. First, we consider the revenue over the time interval ∂t which denotes the duration of one state s_t. Then, recursively we will be able to derive the maximum CSP revenue and hence the optimal price vector. The probabilities to move to any state s_j, where $j = 1, 2, \ldots, k$ given that the current state is s_t at time t can be derived from the state transition matrix Π. Hence, the long-term excepted CSP revenue can be defined as follows,

$$\mathcal{R}_e^*(T) = \underset{\mathbf{P}_e}{lub}[\underbrace{p_e(s_t)x_e(t)\partial t}_{\text{actual revenue at } \partial t} + \underbrace{\sum_{j=1}^{k} \pi(s_j|s_t)\partial t.\mathcal{R}_e^*(T - \partial t)]}_{\text{expected revenue at } T-\partial t} \quad (9)$$

where, the optimal expected CSP revenue $\mathcal{R}_e^*(T)$ consists of the actual revenue during time ∂t, denoted by $p_e(s_t)x_e(t)\partial t$, where $x_e(t)$ is the actual pool size reserved to fulfill the bandwidth demands, and the second term represents the expected revenue from the remaining time $T - \partial t$ depending on the probabilities of the workload distribution according to the next state s_j. The expected pool size at state s_j can be calculated as $x_e(T - \partial t) = n\mu(s_j) + z_{1-\alpha}\sqrt{n}\sigma(s_j)$ following the results from the previous section. Thus, the optimal revenue $\mathcal{R}_e^*(T)$ and the optimal price vector $\mathbf{P}_e^*$ can be calculated by recursively substituting of the equivalent expression of $\mathcal{R}_e^*(T - \partial t)$ into (9), and the boundary conditions $\mathcal{R}_e(0) = 0$ $\forall x_e(t) \in [0, c_e]$.

It is worth noting here that as Markov model is Learnt, a *tâtonnement*-like procedure is used to facilitate the corresponding rate of convergence of the prices. During this process, the optimal prices are calculated based on the state transition matrix until it converges to an equilibrium state when the prices reflect the actual supply and demand. After estimating the optimal long-term revenue $\mathcal{R}_e^*(T)$ through the operating time cycle $[0, T]$, The CSP is able now to compare the revenue of providing the EaaS, and the expected income from selling the extra network capacities to new cloud requests. In addition, the CSP can make the decision of either saving the network capacities for EaaS demands, or selling them.

VI. Performance Evaluation

To evaluate the accuracy of our proposed inter-data centers EaaS approach, two paths are followed, namely through real inter-data centers' traces and simulations. The former is performed based on the Equinix-Chicago traces [29], [33]. The latter involves simulating a set of distributed data centers network, where the links' bandwidth values are uniformly distributed between 250 and 300 Mbps. 30 distributed cloud applications were simulated, where each inter-data center physical link is shared among them. The simulated cloud applications may experience three possible states of operation, which are characterized by three different traffic workload parameters; state(1) $\mu_1 = 10$ Mbps, $\sigma_1^2 = 10$ Mbps, state(2) $\mu_2 = 15$ Mbps, $\sigma_2^2 = 20$ Mbps, and state(3) $\mu_3 = 25$ Mbps, $\sigma_3^2 = 30$ Mbps. The Markov state transition matrix between states was generated randomly from a normal distribution and normalized to satisfy the transition matrix properties. The experiments simulated two weeks of operation.

Estimated workload at different elasticity levels
Figures 2 presents the actual traffic according to Equinix traces and the estimated fluctuation for four different application types; namely HTTP, HTTPS, UDP and RTMP (the value of α was set to 0.5). For simplicity, only the weekdays are considered. The period needed to learn Markov states was six weeks. It is clear that the repeated workload pattern enhances the accuracy of the proposed EaaS model.

To analyze the impact of the elasticity level α on the amount of reserved network resources for the EaaS demands. Fig. 3 depicts the simulated workload fluctuation and the reserved bandwidth at three elasticity levels (i.e., $\alpha = 0.05, 0.225, 0.5$). We can easily see that at high elasticity level with a low probability of unmet bandwidth demands (i.e., low α), the estimated workload and hence the reserved amount of network resources are always more than then the actual workload to prevent violating the EaaS level. However, at a lower elasticity level (i.e., $\alpha = 0.5$), the estimated workload can be close to the actual fluctuation but with a high probability that the actual demand exceeds the estimated one.

Furthermore, Fig. 4 presents the percentage of the met service elasticity demands at different elasticity levels. It can be seen that at high elasticity levels (i.e., $1 - \alpha \sim 1$), the excess demand is met with high probability, however, at medium elasticity levels (i.e., $1-\alpha \sim 0.5$), about 50 percent of

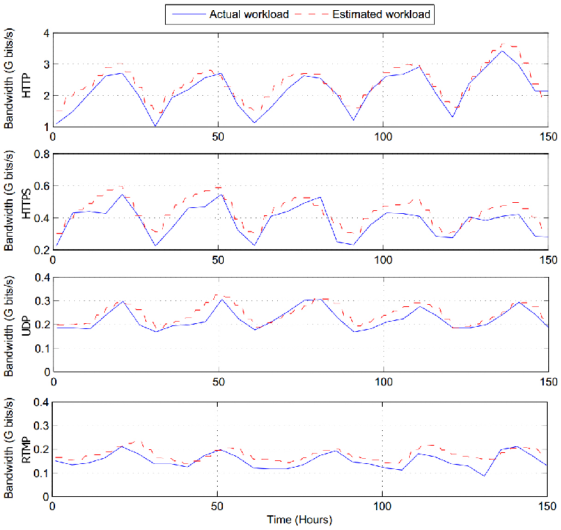

Fig. 2. The actual workload vs the estimated (Equinix traces)

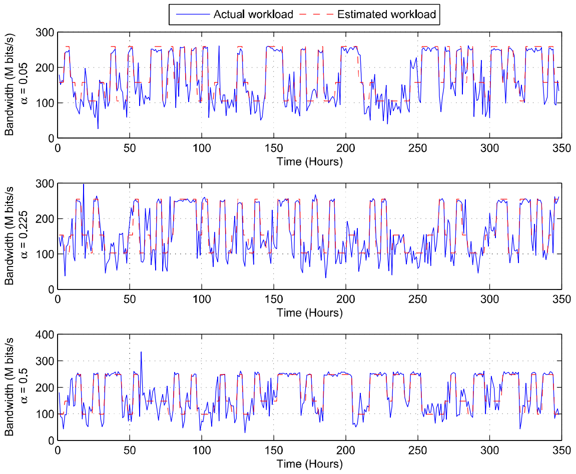

Fig. 3. Inter-data centers workload fluctuation

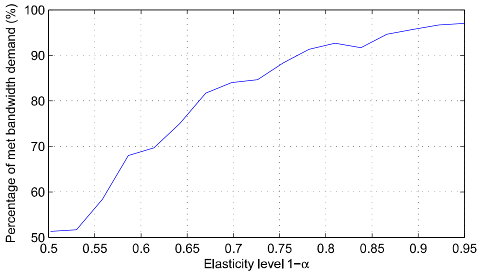

Fig. 4. The percentage of the met bandwidth demands at different EaaS levels

the excess elastic demands are met with the reserved pool size.

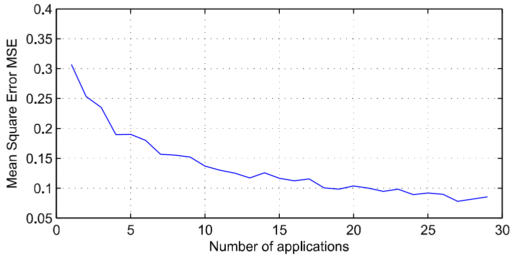

Fig. 5. Elasticity pool reservation accuracy

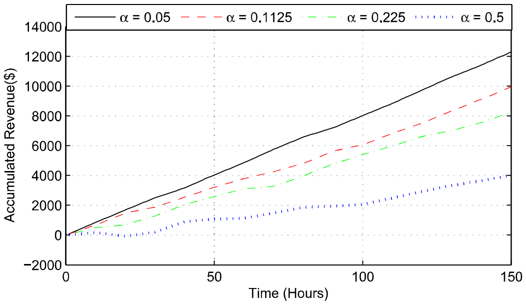

Fig. 6. Accumulated CSP revenue

Estimated traffic workload accuracy

In order to evaluate the accuracy of our proposed approach on deciding the amount of resources needed to be reserved for future demands elasticity, we have measured the *Mean Square Error* MSE when different number of applications, n, are being hosted as shown in section IV). The MSE is defined as the expected sum of squared differences between the estimated and the corresponding actual fluctuations. As can be seen from Fig. 5, when the number of the applications sharing the same class per link increases, the MSE becomes closer to zero. This demonstrates the scalability of this model in cloud environments. It also can be seen that when the number of the cloud applications varies from 10 to 30, the MSE values increase from 0.14 to 0.04 which still are acceptable values.

Fig 6 plots the CSP accumulated revenue, which is defined as the profit from selling the EaaS at four elasticity levels (i.e., $\alpha = 0.05, 0.1125, 0.225, 0.5$) minus the cost of reserving extra resources. The plot shows that the CSP accumulated revenue obtained from high elasticity class is higher since more bandwidth demands are met at a higher price.

VII. Conclusion and Future Work

In this paper, we have presented a novel approach which enables the CSP to provide inter-data centers EaaS at differentiated levels for geographically distributed cloud applications. The approach accurately estimates the amount of resources needed to be reserved to fulfill the future elastic inter-data centers workload fluctuation. Moreover, we have proposed a corresponding EaaS dynamic pricing model that aims at maximizing the CSP expected long-term revenue. Experimental results have shown the accuracy of the model, and the increase in the CSPs profit. In future work, we plan to consider elastic network management for several classes of coexistent heterogeneous distributed cloud applications.

References

[1] Q. Zhang, Q. Zhu, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic service placement in geographically distributed clouds," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 12, pp. 762–772, December 2013.

[2] A. Mahimkar, A. Chiu, R. Doverspike, M. D. Feuer, P. Magill, E. Mavrogiorgis, J. Pastor, S. L. Woodward, and J. Yates, "Bandwidth on demand for inter-data center communication," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, ser. HotNets-X. New York, NY, USA: ACM, 2011, pp. 24:1–24:6.

[3] A. Ghosh, S. Ha, E. Crabbe, and J. Rexford, "Scalable multi-class traffic management in data center backbone networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 12, pp. 2673–2684, 2013.

[4] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *Proceedings of the IEEE INFOCOM 2013, Turin, Italy*, 2013, pp. 854–862.

[5] N. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Comput. Netw.*, vol. 54, no. 5, pp. 862–876, Apr. 2010.

[6] N. Bitar, S. Gringeri, and T. Xia, "Technologies and protocols for data center and cloud networking," *Communications Magazine, IEEE*, vol. 51, no. 9, pp. 24–31, September 2013.

[7] J. Rao, Y. Wei, J. Gong, and C.-Z. Xu, "QoS guarantees and service differentiation for dynamic cloud applications," *Network and Service Management, IEEE Transactions on*, vol. 10, no. 1, pp. 43–55, March 2013.

[8] K. Li, J. Wu, and A. Blaisse, "Elasticity-aware virtual machine placement for cloud datacenters," in *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*, Nov 2013, pp. 99–107.

[9] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*, Oct 2010, pp. 9–16.

[10] A. Shieh, S. Kandula, A. Greenberg, C. Kim, and B. Saha, "Sharing the data center network," in *Proceedings of the 8th USENIX conference on Networked systems design and implementation*, ser. NSDI'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 23–23.

[11] V. T. Lam, S. Radhakrishnan, R. Pan, A. Vahdat, and G. Varghese, "Netshare and stochastic netshare: predictable bandwidth allocation for data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 3, pp. 5–11, Jun. 2012.

[12] D. Niu, C. Feng, and B. Li, "Pricing cloud bandwidth reservations under demand uncertainty," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 151–162, Jun. 2012.

[13] D. D. Mon and M. Gurusamy, "Towards flexible guarantees in clouds: Adaptive bandwidth allocation and pricing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 99, p. 1, 2014.

[14] L. Yu and H. Shen, "Bandwidth guarantee under demand uncertainty in multi-tenant clouds," in *Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on*, June 2014, pp. 258–267.

[15] T. Ghazar and N. Samaan, "Pricing utility-based virtual networks," *Network and Service Management, IEEE Transactions on*, vol. 10, no. 2, pp. 119–132, June 2013.

[16] G. Carella, T. Magedanz, K. Campowsky, and F. Schreiner, "Elasticity as a service for federated cloud testbeds," in *Communications Workshops (ICC), 2013 IEEE International Conference on*, June 2013, pp. 256–260.

[17] S. Pacheco-Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr, and S. Dawson, "Markovian workload characterization for qos prediction in the cloud," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, July 2011, pp. 147–154.

[18] A. Beloglazov and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 7, pp. 1366–1379, July 2013.

[19] Amazon web services AWS. Accessed: 2014-06-12. [Online]. Available: http://aws.amazon.com/

[20] H. C. Lim, S. Babu, J. S. Chase, and S. S. Parekh, "Automated control in cloud computing: Challenges and opportunities," in *Proceedings of the 1st Workshop on Automated Control for Datacenters and Clouds*, ser. ACDC '09. New York, NY, USA: ACM, 2009, pp. 13–18.

[21] W. Dawoud, I. Takouna, and C. Meinel, "Elastic vm for cloud resources provisioning optimization," in *Advances in Computing and Communications*, ser. Communications in Computer and Information Science, A. Abraham, J. Lloret Mauri, J. Buford, J. Suzuki, and S. Thampi, Eds. Springer Berlin Heidelberg, 2011, vol. 190, pp. 431–445.

[22] N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in *Proceedings of the 2011 IEEE 4th International Conference on Cloud Computing*, ser. CLOUD '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 500–507.

[23] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, "A cost-aware elasticity provisioning system for the cloud," in *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, June 2011, pp. 559–570.

[24] Scryer: Netflix. Accessed: 2013-10-06. [Online]. Available: http://techblog.netflix.com/2013/11/scryer-netflixs-predictive-auto-scaling.html

[25] 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 *Proceedings of the 6th International COnference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 15:1–15:12.

[26] T. Benson, A. Akella, A. Shaikh, and S. Sahu, "Cloudnaas: a cloud networking platform for enterprise applications," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*, ser. SOCC '11. New York, NY, USA: ACM, 2011, pp. 8:1–8:13.

[27] D. Divakaran and M. Gurusamy, "Probabilistic-bandwidth guarantees with pricing in data-center networks," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 3716–3720.

[28] W. D. Mulia, N. Sehgal, S. Sohoni, J. M. Acken, C. Lucas Stanberry, and D. J. Fritz, "Cloud workload characterization." *IETE Technical Review*, vol. 30, no. 5, 2013.

[29] Equinix inter-data centers traces. Accessed: 2015-01-05. [Online]. Available: www.caida.org/

[30] D. Niu, Z. Liu, B. Li, and S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 421–425.

[31] A. Papoulis, *Probability, random variables, and stochastic processes*. McGraw-Hill, 1984.

[32] D. P. Bertsekas, *Dynamic programming and optimal control*. Vol. 1. No. 2. Belmont, MA: Athena Scientific, 1995.

[33] Equinix data centers. Accessed: 2015-01-05. [Online]. Available: www.equinix.com/