Service Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Learning Curves and Stochastic Models for Pricing and Provisioning Cloud Computing Services

Amit Gera, Cathy H. Xia,

To cite this article:
Amit Gera, Cathy H. Xia, (2011) Learning Curves and Stochastic Models for Pricing and Provisioning Cloud Computing Services. Service Science 3(1):99-109. https://doi.org/10.1287/serv.3.1.99

Service Science 3(1), pp. 99-109, © 2011 SSG

Learning Curves and Stochastic Models for Pricing and Provisioning Cloud Computing Services

Amit Gera, Cathy H. Xia
Dept. of Integrated Systems Engineering
Ohio State University,
Columbus, OH 43210
{gera.2, xia.52}@osu.edu

The paradigm of cloud computing has started a new era of service computing. While there are many research efforts on developing enabling technologies for cloud computing, few focuses on how to strategically set price and capacity and what key components are leading to success in this emerging market. In this paper, we present quantitative modeling and optimization approaches for assisting such decisions in cloud computing services. We first show that learning curve models can be helpful to capture the providers' cost reduction with economy of scale. Such models also help understand the potential market of cloud services and explain quantitatively why cloud computing is most attractive to small and medium businesses. We then present a stochastic model and a revenue management formulation to address the pricing and resource provisioning decisions for the cloud service providers. The approach enables the cloud service provider a quantitative framework to obtain management solutions and to learn and react to the critical parameters in the operation management process by gaining useful business insights.

*Key words***:** cloud computing; learning curves; stochastic modeling; pricing; provisioning
History: Received Aug. 18, 2010; Received in revised form Aug. 30, 2010; Accepted Sep. 13, 2010; Online first publication Nov. 12, 2010

1. Introduction

Fueled by the dramatic growth in connected devices, real-time data streams, and the adoption of service oriented architectures and Web 2.0 applications, cloud computing is rapidly gaining momentum as an alternative to traditional IT. Cloud computing represents an emerging service paradigm in which large pools of systems are linked together to provide IT services on behalf of clients (see, e.g. Hayes (2008). The key features of cloud computing are on-demand availability and pay per use policy. On-demand availability gives a customer the flexibility to increase or decrease capacity on the fly to meet its demand. It allows companies to trade fixed costs for variable usage based costs. Often the fixed costs of infrastructure are large and prohibitive to small to medium size businesses. Cloud computing also promises to reduce the human administration and development costs.

In this emerging cloud computing services marketplace, Amazon, Google, IBM, Microsoft and Salesforce.com are leading the way with innovative service offerings. Several pricing and capacity planning strategies have emerged amongst cloud computing service providers. For example, it has been reported that Salesforce.com uses 10 times fewer servers than Amazon to support the same number of clients (see techcrunch.com). While there are many research efforts on developing enabling technologies (such as virtualization, standardization, parallel computing, and open source software) for cloud computing, few focus on what leads to these various different pricing and provisioning strategies, how to strategically set the proper pricing and capacity, and what key components are leading to success in this emerging yet highly competitive market?

According to IDC's recent Cloud services user survey (see IDC) of 244 IT executives/CIOs and their line-of-business colleagues, the top two attributes users want from cloud services providers are competitive pricing and offering performance-level assurances. While the value delivered by cloud services are clear to the small and medium businesses, the cloud provider may have trouble pricing and operating profitably. To attract business, the cloud provider must offer a price visibly lower than what the customer believes it is spending on its in-house management of IT. The provider must also provision sufficient capacity to satisfy the service level agreement with the customers. Hence, cost reduction and revenue generation are critical to the success of a cloud service provider. Consequently, strategic solutions for pricing, provisioning, and delivering quality of service are the determinant components of the business process in order to win in the competitive emerging market.

In the literature, pricing and revenue management problems have received much attention over the last decade in many areas, ranging from the airline yield management to hotel industry to retail stores (see Petruzzi and Dada (1999)). There exists a vast literature on revenue management, and interested readers can be referred to, e.g. Talluri and van Ryzin (2004) for detailed reviews.

In the context of pricing and revenue management of utility computing services in the on-demand paradigm, there has been some recent work. Gurnani and Karlapalem (2001) studies the economic viability of pay-per-use licensing of packaged software over the internet by using a monopoly pricing model. Paleologo (2004) accounted for the impact of uncertainty in the decision process in the pricing of on-demand computing services and proposed a novel methodology ``price-at-risk." Dube et al. (2007) studies the competitive equilibrium in the computing service market when competition is on both pricing and quality of service, and customers further face the outsourcing or in-house hosting decisions.

Many studies also address the queueing effects in the problem domain. Mendelson (1985) studied the effect of congestion due to queueing effects and users' related costs on pricing and capacity decisions at a computing center with a single processor. Mieghem (2000) studied the optimal prices and service quality grades of a service provider facing heterogeneous utlity-maximizing customers in queueing models under either full information or asymmetric information. Hampshire et al. (2002) used queueing models to solve QoS provisioning and pricing problems under Erlang loss queueing models.

The above literature review is by no means complete. However, the pricing and resource provisioning problem remains challenging in the cloud computing paradigm in a number of ways. Often, a cloud service provider offers a menu of service templates or configurations from which companies choose. The cloud provides at least three distinct resources whose combination makes the final products on the menu. The templates are composed of various amounts of three resources: CPU, memory and storage. Clients may choose to purchase any combination of these templates. The cloud provider gives customers the guarantee that their template will be available and uncompromised by other customers' usage.

The resource provisioning is therefore of multiple dimensions (CPU, memory, bandwidth, storage). Different customers may have different requirements on each of these dimensions, making the provisioning problem much more difficult. The Cloud flexibility paradigm allows customers to increase and decrease capacity on the fly adds further complications on the provisioning task. Lastly, as an emerging market, the elasticity of demand (to price and quality of service) in cloud computing is not well understood. Due to lack of historical data, demand will be initially difficult to forecast. Also, it will take time for economies of scale to be realized.

In this paper, we outline quantitative modeling and optimization approaches that might lead to viable solutions to these challenges. In section 2, we propose to use a learning curve model to capture the economy of scale in cloud computing services. Such models can also be used to study customers' cloud adoption decisions and help understand quantitatively why cloud computing is most attractive to small and medium businesses. In section 3, we introduce a stochastic model that captures the dynamics in cloud computing services. We then study the steady-state behavior of such systems and present a revenue management formulation that helps to find the optimal pricing and provisioning solutions. We further demonstrate the approach via numerical examples and provide management solutions that lead to useful business insights.

2. Learning Curve Model for Cloud Computing

In cloud computing, a significant amount of expertise and learning is involved in operating large distributed data centers efficiently. In this section, we present a learning curve model that captures the production cost reduction with economy of scale in cloud computing. We further use the model to study customers' cloud adoption decisions and show quantitatively why cloud computing is most attractive to small and medium businesses.

2.1 Learning Curves

Learning curves, first proposed in Wright (1936), are used to determine the reduction in costs per unit as experience is gained while producing products with mature and immature process technologies (Blancett 2002). It is a useful model to explain the economy of scale and has been successfully applied to a wide variety of industries ranging from aerospace to auto and ship building to semiconductor manufacturing (Russell and Taylor 2003).

The learning curve model assumes that as the number of production units are doubled the marginal cost of production decreases by a fixed factor (or percentage). One minus this factor is called learning factor (or learning coefficient). Mathematically, it boils down to a widely accepted power law model, where the marginal cost of producing the x-th unit, denote by $c(x)$, is given by

$$c(x) = Kx^{\log_2 \alpha} \tag{1}$$

where K denotes the cost of the first unit, and $\alpha \in (0,1)$ is the learning factor.

The total cost of producing x units, denoted by $C(x)$, can then be obtained by integrating the above marginal cost function for *0* to x, which gives

$$C(x) = \frac{Kx^{1+\log_2 \alpha}}{1+\log_2 \alpha} \tag{2}$$

In cloud computing, companies like Amazon have developed extensive in-house knowledge which they leverage to sell services at a margin. It is estimated that Amazon currently is running 30,000 EC2 instances, and the cost per instance is currently $.08/hr. Their original cost per EC2 instance was estimated to be $2.20/hr. This represents an 80% learning factor. It has been reported [3] that for a typical data center, as the total number of servers in house doubles, the marginal cost of deploying and maintaining each server decreases 10-25%, thus the learning factors are typically within the range (0.75, 0.9).

2.2 Customer Cloud Adoption Decision

Suppose the production cost function $C_s(x)$ of the service provider is given by:

$$C_s(x) = \frac{K_s x^{1+log_2 \alpha_s}}{1+log_2 \alpha_s} \tag{3}$$

where α_s denotes the learning factor of the service provider, and K_s is the production cost of 1st unit incurred to the service provider. The total cost on the learning curve divided by the number of production units, then gives the unit price that the service provider would charge to its customers.

Assume customers can meet their computing requirements in two ways: either in-house or outsourced via cloud service. Each customer makes his/her decision of adopting cloud service or not based on his cost tradeoffs. The cost for the customer to meet x units of service *in-house* is given by

$$C_c(x) = \frac{K_c x^{1+\log_2 \alpha_c}}{1+\log_2 \alpha_c} \tag{4}$$

where α_c is the learning factor of the customer, K_c is the production cost of 1st unit by the customer. Hence, if the cost of doing in-house is less, then customer will not opt for cloud computing. So, hence, if a customer demands x units of service, then he will only accept cloud computing if

$$C_c(x) > x\pi \tag{5}$$

where π is the price per unit time for a unit of service charged by the (cloud) service provider. Solving above inequality gives:

$$x < \left(\frac{K_c}{(1+\log_2 \alpha_c)\pi} \right)^{\left(\frac{-1}{\log_2 \alpha_c}\right)} =: \eta(\pi) \tag{6}$$

The learning curve model explains naturally why cloud service offering is mostly attractive to small and medium businesses. Figure 1 plots the total cost for two companies with different learning curve factors. Although both companies have the same cost of first unit as $1500, the total cost of production for red ($\alpha = 0.75$) is much lower than for blue ($\alpha = 0.8$). Consequently, the red company can sell units at much lower cost than blue company.

Figure 2 shows the cost comparison of the two service options for a company, where the concave line corresponds to the in-house cost function based on the learning curve model while the straight line corresponds to the linear cost using cloud service. We see that it would be better off by adopting cloud service if its demand is smaller than $\eta(\pi)$, thus cloud service is generally attractive to small to medium businesses. In order to get under the service provider's unit price, one would have to be deploying tens of thousands servers in house.

From the service provider's point of view, $\eta(\pi)$ indicates the potential market share it can capture under a given unit price π. Figures 3 and 4 show how $\eta(\pi)$ changes as a function of π for different values of parameters α_c and K_c respectively. From the figures, we see that the demand is sensitive to both the learning factor and the price. The harder it is for customers to learn, the more will come to cloud services. On the other hand, inappropriate

pricing can definitely drive away demand from the cloud service provider. Similarly, the demand is sensitive to the first unit production cost K_c. The higher the cost of first unit, the more will come to cloud service. Thus, customers with less expertise in computing services would tend to find cloud service attractive.

Figure 1 Learning Curves under Different Learning Factors

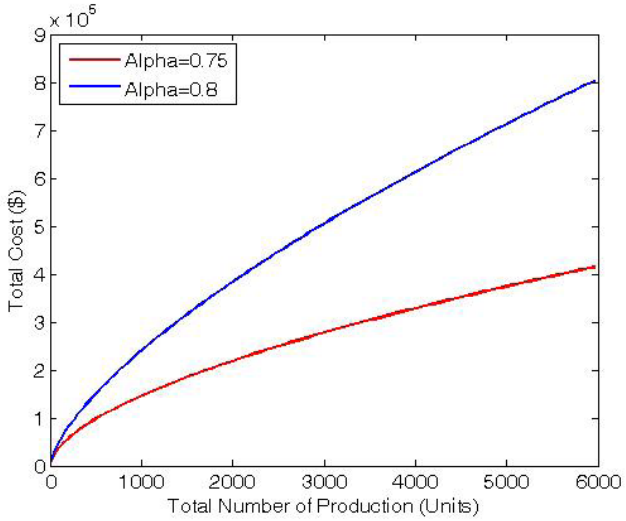
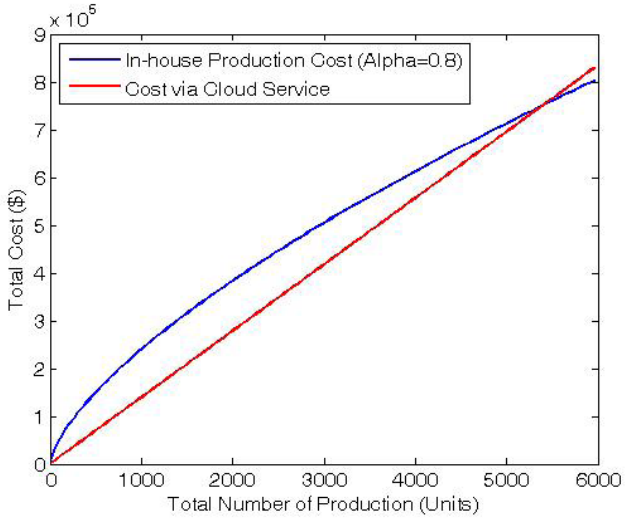


Figure 2 Cost Comparison: In-house vs. Cloud Service



3. Joint Pricing and Resource Provisioning

In this section, we study the problem of pricing and resource provisioning from the service provider's point of view. We first introduce a stochastic model that captures the flexibility of allowing customers to upgrade and downgrade their services dynamically. We then study the steady-state behavior of such systems and present a revenue management formulation that helps to find the optimal pricing and provisioning solutions. We further demonstrate the approach via numerical examples and provide management solutions that lead to useful business insights.

3.1 The Model

Consider a single cloud service provider who provides J classes of service templates. For simplicity, we consider the case where all J classes are defined as multiple units of some base instance. A class j service template consists of m_j units of base instance, with $j=1, \ldots, J$. Assume the service classes are ordered such that $m_1 < m_2 < \ldots < m_J$. For example, a base instance at Amazon EC2 is comprised of 1.7 GB of memory, 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit) and 160 GB of instance storage.

Assume customers desiring class *i* services arrive according to a Poisson process with the rate λ_j (with λ_j to be specified, and it depends on both the price and the desired service quality).

Figure 3 η(π) for varying α's

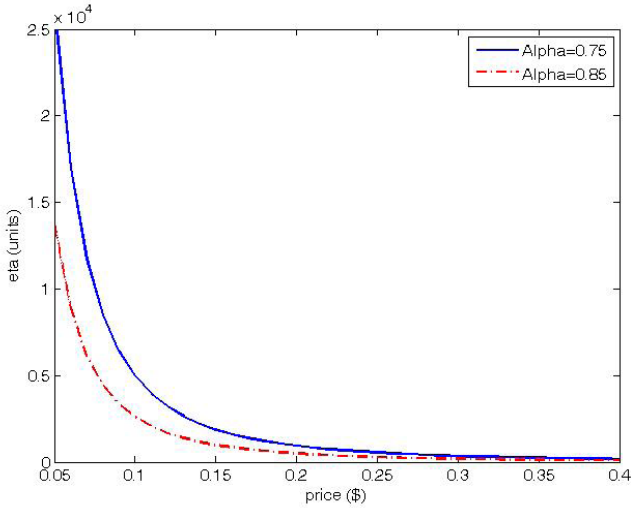
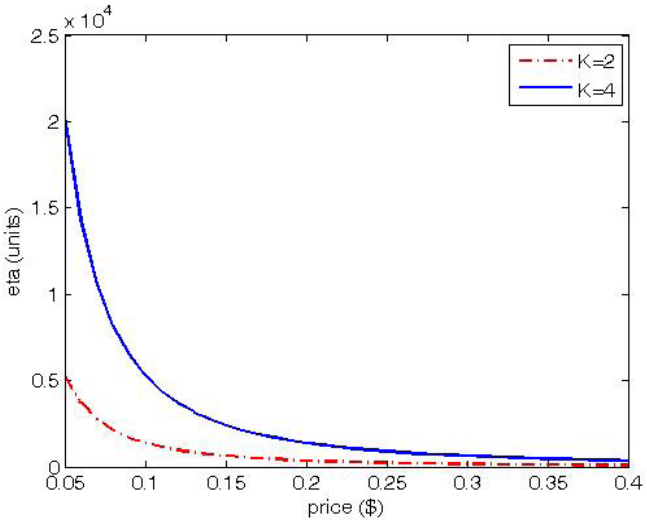


Figure 4 η(π) for varying K's



To capture the flexibility of allowing customers to increase or decrease resource requirements on the fly in cloud computing services, we consider the following probabilistic model. Assume that, after holding the service template *j* for an exponential amount of time with mean $1/\mu_j$, a customer upgrades/downgrades to service template *k* with probability p_{jk} and terminate the service with probability $1-\sum_{k=1}^{J} p_{jk}$. The matrix $P = [p_{jk},\ 1 \le j,\ k \le J\ ]$ is transient so that every customer eventually leaves, thus *I-P* is invertible. The holding times for class *i* templates are i.i.d. random variables, independent of the arrival processes and other classes.

We assume the cloud service provider operates in a market with imperfect competition. This is suitable for cloud computing as it is a new market and the provider may hold a temporary monopoly, or the market may allow for product differentiation. In the monopoly or new product case under imperfect competition, the provider can influence demand by varying its price. We assume the demand rate for class *j* service templates depends only on the price π_j through a non-increasing function $\lambda_j(\pi_j)$ where π_j denotes the price charged for providing service template *j* per unit time. An alternative view of such demand function is to assume that customers (of class *j*) have i.i.d. reservation price with distribution F_j, where the reservation price represents the maximum price that a customer is

willing to pay. Therefore, the expected demand rate at price π_j is $a_j F_j(\pi_j)$, where a_j is the potential demand rate of class j customers before the price is announced.

The above reservation function can be derived using the learning curve model we discussed earlier. Since different customers tend to have different capacity requirements and learning capabilities, it requires extra efforts to understand the distribution of these variables. One can rely on marketing analysis, such as panel studies, survey research to obtain some sort of estimation on these parameters. In this paper, we simply assume that $\lambda_j(\pi_j)$ is given as a function of π_j and focus on the revenue management problem.

Cloud service providers are compelled to deliver a quality of service (QoS) defined by the underlying service level agreement. Typical service level agreements are termed such that a customer should have access to its desired service with probability $1-\epsilon$. We consider an availability guarantee where the cloud provider guarantees that a customer should have access to its desired capacity with probability $1-\epsilon$, uncompromised by other customers' usage at all times.

The service provider needs to choose proper prices for each service class, $\pi = (\pi_1, ..., \pi_J)$. The service provider also needs to provision sufficient capacity so as to meet the availability guarantee.

3.2 Steady-State Behavior

Different from many covariance-based modeling approaches, the Partial Least Squares approach to Structural Mathematically, one can view the customer flexibility model as a service with phase type distribution. That is, customers (of all types) arrive according to a Poisson process with mean arrival rate $\lambda=\lambda_1 + ... + \lambda_J$. Upon arrival, each customer follows a finite state continuous time Markov chain (CTMC) on the state space $\{0, 1, ..., J\}$, where states $1, ..., J$ are transient and state 0 is absorbing. State i represents a customer using service template i, $i=1, ..., J$. Once reaching the absorbing state 0, the customer leaves the system. The initial distribution is $(0, \alpha)$, where $\alpha=(\alpha_1, ..., \alpha_1)$ and $\alpha_i = \frac{\lambda_i}{\lambda}$. The CTMC has infinitesimal generator as $\begin{pmatrix} 0 & \vec{0} \\ \vec{t} & S \end{pmatrix}$, where $S_{jk}=\mu_j p_{jk}$ for $j,k=1, ..., J$, and $t=-S\mathbf{1}$. Here $\mathbf{1}$ represents the unity (column) vector with all entries being 1.

Let X denote the sojourn time of a customer until absorption. Then X follows the so-called phase-type distribution $PH(\alpha, S)$ (see e.g. Latouche and Ramaswami (1999)). The mean sojourn time $E[X]$ is given by:

$$E[X] = -\alpha S^{-1}\mathbf{1}.$$

Now view the above sojourn time as the service time for the cloud provider to serve each customer until he leaves the system. If the service provider has an infinite amount of resource available, the number of customers in system at any time t is the same as the number of customers of a $M/G/\infty$ queue where the service time for each customer is X with phase-distribution $PH(\alpha, S)$.

Denote N the total number of customers in the system in steady state, and N_j the number of customers in state j, $j=1, ..., J$. Based on the theory on $M/G/\infty$ (see e.g Gross and Harris (1985)), the total number of customers in system is Poisson distributed with mean $\rho := \lambda E[X]$. Furthermore, N_j, the number of customers of type i in system is also Poisson distributed with mean ρ_j, where $\rho_j = \lambda(-\alpha S^{-1})(j)$, and $\rho = \sum_{j=1}^{J} \rho_j$. Note that $N_1, ..., N_J$ are independent Poisson random variables. Consequently, in steady-state, the probability that the system is in state $(n_1, ..., n_J)$ is given by:

$$\pi(n_1, ..., n_J) = \prod_{j=1}^{J} e^{-\rho_j} \frac{\rho_j^{n_j}}{n_j!} = e^{-\rho} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \tag{7}$$

3.3 Resource Provisioning

The service provider needs to provision sufficient in house capacity so as to meet the service availability guarantee $1-\epsilon$. We next decide the total capacity B based on the steady-state behavior of the system.

In steady state, *N*, the total number of customers in system is Poisson distributed with parameter ρ . Label the *N* customers randomly and let Y_i be the capacity requirement (in terms of number of base instances) of the *i*-th customer, *i=1, ..., N*. Since a randomly selected customer is of type *j* with probability $q_j = \rho_j / \rho$, thus

$$q_j = \rho Y_i = m_j, \text{ with probability } q_j, \quad j = 1, ..., J_{\cdot j} / \rho$$

Hence, the total capacity requirement to the system is given by $\sum_{i=1}^{N} Y_i$. The goal of capacity planning is to choose *B* large enough such that

$$P\{\sum_{i=1}^{N} Y_i > B\} \le \epsilon \tag{8}$$

Notice that $\sum_{i=1}^{N} Y_i$ is a compound Poisson (CP) random variable with mean $\rho E[Y]$. For convenience, we denote such a compound Poisson random variable by CP(ρ, Y).

To decide the proper level of capacity *B*, we utilize a result derived in Madiman and Kontoyiannis (2005), which gives an upper bound on the tail distribution of a compound Poisson random variable. We will state the theorem in Madiman and Kontoyiannis (2005), first and then derive a closed form expression on how to select *B* to achieve availability guarantees $1-\epsilon$.

Theorem 1: *Suppose U has a distribution CP(ν , W), where the base distribution (of W) has finite support*,

$$m = max(i \in Z_+ : W_i > 0) < \infty$$

Let $f : Z_+ \to \mathbb{R}$ *be S-Lipschitz, i.e.* $| Df(x) | = | f(x+1) - f(x) | \le S$ *for all x. Then if* $t \in \left[0, (3\nu m^2)/2\right]$

$$P[f(U) - E[f(U)] \ge t] \le exp\left(-\frac{t^2}{3\nu m^3}\right) \tag{9}$$

Now consider $U = \sum_{i=1}^{N} Y_i$ which is a compound Poisson random variable CP(ρ, Y). Note that the Y_i's has a finite support $m=max\{m_j: j=1, ..., J\}$. Choose the identical function $f(x)=x$, which is clearly S-Lipschitz. We can then apply the above theorem to choose a *B* that satisfies (8).

It suffices to let $B = E[\sum_{i=1}^{N} Y_i] + t = \rho \sum_{j=1}^{J} q_j m_j + t$, where *t* is chosen such that

$$ln(\epsilon) = -\frac{t^2}{3\rho m^3}$$

We then have

$$t = [-3\rho m^3 ln(\epsilon)]^{1/2} \tag{10}$$

But (9) is only valid if *t* is in the range given in the theorem. Thus, we need to make sure

$$[-3\rho m^3 ln(\epsilon)]^{1/2} \le \frac{3\rho m^2}{2}$$

Solving the above inequality will give us,

$$\rho m \ge -\frac{4}{3} ln(\epsilon) \tag{11}$$

Condition (11) can be easily satisfied in the cloud computing setting. For example, to achieve service availability 99.95% (thus ϵ = 0.05%) the right hand side of (11) will be in the order of 10. Since ρ and *m* are typically large in cloud services, (e.g. the mean load ρ is typically in the order of hundreds or thousands, and *m* can be in the order of 10's or 100's), thus condition (11) can be easily satisfied.

Therefore, in order to provide service availability $1-\epsilon$, it suffices to provision capacity *B* such that

$$B = \rho\sum_{j=1}^{J} q_j m_j + [-3m^3 ln(\epsilon)]^{1/2}\sqrt{\rho} \tag{12}$$

Observe from (12) that the planned capacity *B* consists of two terms. The first term is the expected value of resource requirement in steady state which is linear to the mean load ρ. The second term ensures a safety margin in the order of $\sqrt{\rho}$ to absorb the demand variability where the coefficient depends on the target availability ϵ. We believe such square root staff result can be extended to the more general case where the customer sojourn times are generally distributed.

3.4 The Revenue Management Problem

We are now ready to formulate the joint pricing and resource provisioning problem as a revenue management problem for the cloud service provider from a steady-state point of view.

The profit earned by the service provider is equal to the revenue minus the cost of providing the service. Under a committed service availability $1-\epsilon$, revenue is equal to $\sum_{j=1}^{J}\pi_j\rho_j(\pi_j)(1-\epsilon)$. This is the summation of the price charged per hour per customer for each service class multiplied by the average number of customers in that class thinned by a ϵ fraction of demand loss.

In order to ensure service availability $1-\epsilon$, the provider will provision a total capacity *B* (in terms of number of base instance) given by (12). The cost of providing services at capacity level *B* is given by the provider's learning curve $C_s(B)$ as defined in (3).We then have the following optimization problem:

$$\max_{\pi,B}\sum_{j\in J}\pi_j\rho_j(\pi_j)(1-\epsilon) - C_s(B) \tag{13}$$

Note that capacity *B* also depends on the prices π indirectly through ρ_j's.

3.5 Numerical Results

In this section we demonstrate through a small example how optimal prices and resource provision solutions can be obtained via the optimization framework as formulated in (13). Although the example may appear to be a bit simple, the solution provides interesting insights into the behavior of the optimal policy.

Consider a cloud service provider that offers three different types of templates: basic, medium and large. A basic service template consists of one unit of the base instance, a medium template consists of 4 units of base instance, while a large template consists of 8. That is: $m_1 = 2$, $m_2 = 4$ and $m_3 = 8$.

The total arrival rate is 100 customers/hr, where 90% of the arrivals are interested in basic templates, 8% are interested in medium size templates, and 2% are interested in large. Upon entering the system, each customer holds its desired service template for an exponential amount of time with mean 20 hours, and then upgrades and downgrades his service probabilistically according to transition matrix *P*, where $P = \begin{pmatrix} 0 & 0.45 & 0.45 \\ 0.45 & 0 & 0.45 \\ 0.45 & 0.45 & 0 \end{pmatrix}$.

For simplicity, we assume the provider has decided to just set the base price π for using one unit of the base instance per unit time. Hence, $\pi_1 = \pi$, $\pi_2 = 4\pi$, and $\pi_3 = 8\pi$. Assume further that customers' hold reservation price against the base price, and the reservation price is exponentially distributed with parameter γ. This will lead to a simple price function

$$\lambda(\pi) = \lambda e^{-\beta\pi}$$

where $\beta > 0$ is price sensitivity parameter. Such exponential demand-to-price elasticity function has been widely used in many previous studies, see, e.g. Gallego and van Ryzin (1994). In this example, we assume $\beta = 0.8$.

We are varying π from \$0.005 to \$1.25 per instance per hour. For example, Amazon's price for the small instance is \$0.1 per instance per hour EC2. We calculate the resulting $\lambda(\pi)$ and then calculate λ_i for each service class according to Poisson splitting.

The production cost of the service provider for provisioning a total B units of capacity, $C_s(B)$, is given by the learning curve model (3). Using Amazon (EC2) as the example, we assume the production cost of first unit is $K_s=2/hr$, and the target service availability is set to 99.95%. We will show the profit under different learning coefficient by varying the learning factor from 75% - 85%.

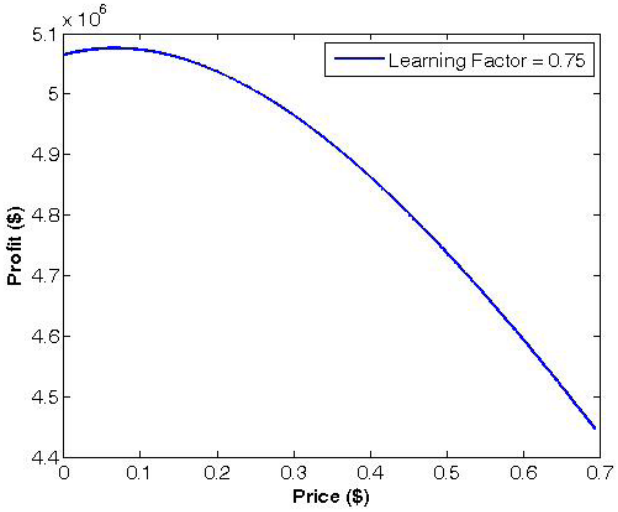
Figure (5) plots the variation of profit vs. price charged under different factors. Observe that with a learning factor $\alpha = 75\%$, the optimal unit price π is less than \$0.10/hr, and the total (monthly) profit is around $\$5.1\times10^6$. With a learning factor $\alpha = 80\%, \pi^* \approx \$0.2\$ / hr$, where the total profit dropped to $\$3.65\times10^6$. When the learning factor is set to $\alpha = 85\%$, it is no longer profitable at all unit prices ranging from 0 to \$1.4/hr.

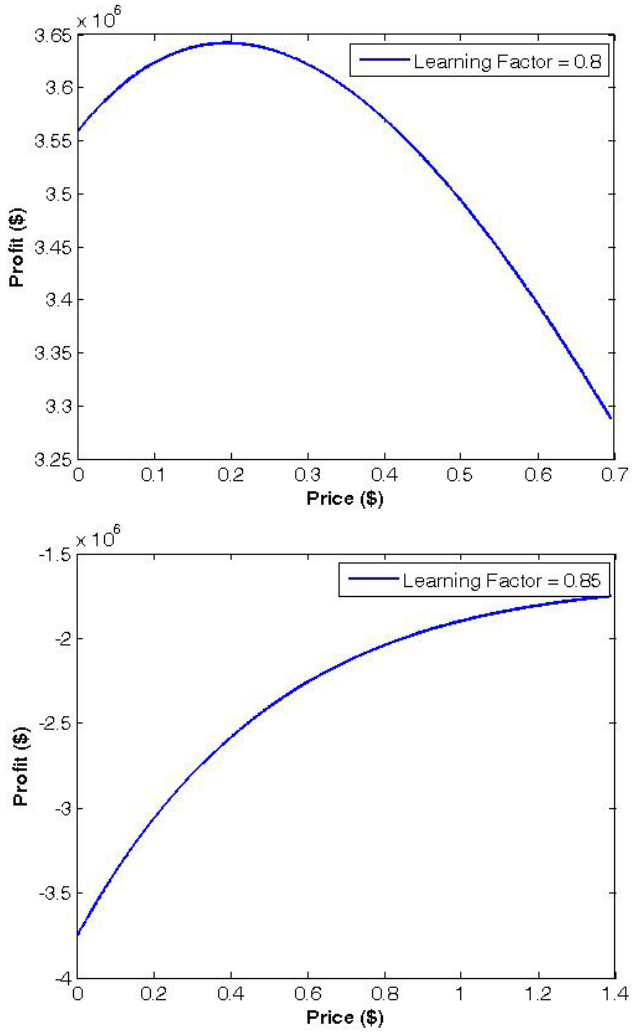
We see that the lower the learning factor (i.e. the faster the service provider can cut down its production cost), the lower the optimal price the provider can offer, and the more profitable it will be. Since a lower price will attract more demands, which yields a higher total profit.

4. Conclusion

In summary, we have presented quantitative modeling and optimization approaches for managing cloud computing services. We showed that learning curve models can be useful to model the providers' cost reduction with economy of scale. Such models also help understand customers' cloud adoption decisions and explain quantitatively why cloud computing is most attractive to small and medium businesses. We further showed that stochastic models and revenue management formulation can help the pricing and resource provisioning decisions for the cloud service providers. The approach enables the cloud service provider a quantitative framework to obtain management solutions and to learn and react to the critical parameters in the operation management process by gaining useful business insights.

Figure 5 Profit vs. Price under Different Learning Factors





References

Amazon Elastic Compute Cloud, http://aws.amazon.com/ec2/.

IDC Enterprise Panel, IT Cloud Services User Survey, pt.3: What Users Want From Cloud Services Providers, August 2008, available at http://blogs.idc.com/ie/?p=213

Dube, P., Z. Liu, L. Wynter, and C. Xia. Competitive equilibrium in ecommerce: Pricing and outsourcing.Computers & Operations Research, 34:3541–3559, 2007

Gallego, G. and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Management Science, 40(8):999–1020, Aug. 1994

Gross, D. and C. Harris. Fundamentals of Queueing Theory. John Wiley & Sons Inc.: New York, 1985

Gurnani, H. and K. Karlapalem. Optimal pricing strategies for internet-based software dissemination. Journal of the Operational Research Society, 52(1):64–70, 2001

Hampshire, R.C., W. A. Massey, D. Mitra, and Q. Wang. Provisioning of bandwidth sharing and exchange. In Telecommunications Network Design and Economics and Management: Selected Proceedings of the 6th INFORMS Telecommunications Conferences (Boca Raton, FL, pages 207–226. Kluwer Academic Publishers, Boston/Dordrecht/London, 2002

Hayes, R. Cloud computing. Communications of the ACM, 51:9–11, 2008

Latouche, G. and V. Ramaswami. Introduction to Matrix Analytic Methods in Stochastic Modelling. ASA SIAM, 1st edition, 1999

Madiman, M. and I. Kontoyiannis. Concentration and relative entropy for compound poisson distributions. In Proceedings of the 2005 IEEE International Symposium on Information Theory (Adelaide, Australia), September 2005

Mendelson, H. Pricing computer services: Queueing effects. Communications of the ACM, 28:312–321, 1985

Mieghem, J.V. Price and service discrimination in queuing systems: Incentive compatibility of gc scheduling. Management Science, 46:1249–1267, 2000

Paleologo, G. A methodology for pricing utility computing services. IBM Systems Journal, 43(1):20–31, 2004

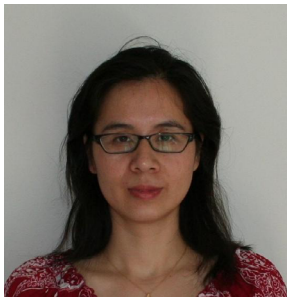
Petruzzi, N. and M. Dada. Pricing and the newsvendor model: a review with extensions. Operations Research, 47:183–194, 1999

Talluri, K. and G. van Ryzin. The Theory and Practice of Revenue Management. Kluwer Academic Publishers, 2004

Techcrunch.com. The Efficient Cloud: All Of Salesforce Runs On Only 1,000 Servers, Mar 23, 2009, available at http://www.techcrunch.com

Wright, T. Factors affecting the cost of airplanes. Journal of Aeronautical Science, 4(4):122–128, 1936

Amit Gera is a M.S. student at The Ohio State University in Industrial & Systems Engineering Department with specialization in Operations Research. Prior to this he got his B.Tech from Indian Institute of Technology, Madras. He has 4 years of experience in stochastic modeling and system performance measures. He is currently working on pricing and capacity planning for Cloud Computing service providers. His research interests include stochastic modeling, revenue management and decision analysis. He is an active member of student chapter INFORMS at The Ohio State University.



Cathy H. Xia is an associate professor in the Department of Integrated Systems Engineering at The Ohio State University, with an adjunct appointment from the Department of Computer Science and Engineering. Dr. Xia specializes in performance modeling and analysis, stochastic processes, and distributed management of computing networks and service systems. She received her M.S. in Statistics, and her Ph.D. in Operations Research, both from Stanford University. Before joining OSU, she has worked as a research scientist at IBM T.J. Watson Research Center for many years and has received numerous IBM research and practice awards. Dr. Xia has served NSF panels and a number of conference committees. She is the program chair of INFORMS Midwest 2011, cluster chair on cloud computing for INFORMS Annual Conference 2010, and tutorial chair for ACM Sigmetrics 2008. Dr. Xia is the associate editor of the International Journal of Information Systems in the Service Sector. Dr. Xia has obtained over 10 patents granted. She has published 1 edited book, 1 book chapter, and over 50 papers in peer-reviewed journals and conferences.