

An Economic Model for Maximizing Profit of a Cloud Service Provider

Thanadech Thanakornworakij¹, Raja Nassar¹, Chokchai Box Leangsuksun¹, Mihaela Paun^{1,2}

¹*College of Engineering & Science, Louisiana Tech University*

Ruston, LA 71270, USA

²*National Institute of Research and Development for Biological Science*

Bucharest, Romania

tth010@latech.edu, nassar@latech.edu, box@latech.edu, mpaun@latech.edu

Abstract—For Infrastructure-as-a-Service, Cloud service providers, such as Amazon EC2 and Rackspace, allow users to lease their computing resources over the Internet, and invest their money into developing and maintaining the infrastructure. Hence, maximizing profit, right pricing, and rightsizing are vital elements to their business. To address these issues, we propose in this article an economic model for cloud service providers that can be used to maximize profit based on right pricing and rightsizing in the Cloud data centre. Total cost is a key element in the model and it is analyzed by considering the Total Cost of Ownership (TCO) of the Cloud.

Keywords— A Cloud Economic model, Profit maximization, Cloud TCO, Availability, Penalty, Cost

I. INTRODUCTION

As demand for Cloud Computing is increasing dramatically, new opportunities are created for running a cloud business. To maximize profit, one has to be able to estimate the demand of the market in order to define the price and the number of resources in order to increase revenue and minimize cost. Based on supply and demand [10], if a company sets the price lower than market price, it can, in principle, increase its sales. On the other hand, if the price is set too high, sales may diminish. Hence, a good pricing strategy is important for increasing revenue. A successful pricing strategy can lead to a significant increase in profit [9]. If one can understand the demand of Cloud Computing, one can define the right price to gain market share and increase revenue.

Another factor that impacts profit of the Cloud service provider is cost. To understand the cost of a Cloud service provider, Total cost of ownership (TCO) [22] is used in the analysis. The total cost of ownership analysis considers direct and indirect cost of owning a business over its life span. The direct cost consists of equipment and infrastructure cost, such as server, network, and facility. The indirect cost includes things like power, cooling, and operating license. When one understands the TCO, one can decide on the right size of a cloud system based on demand in order to maximize profit.

In this article, we propose an economic model for a Cloud service provider, based on revenue and cost, by considering computing resources charged to customers over time. A logit model is considered for revenue analysis and the total cost of ownership is used to analyze the cost of a Cloud system. We

also consider a Service Level Agreement (SLA) which guarantees a certain availability level and apply a probabilistic model to calculate the expected penalty cost when a cloud service provider cannot meet the guaranteed level.

II. RELATED WORK

Pricing and economic analysis have been previously applied in grid and Cloud Computing. Abdelkader [1] studied an economic model for resource allocation in grid computing and treated computational and storage resources as interchangeable. Price was determined by demand and supply at the equilibrium point. Dash [5] proposed a self-tuned economic model for query service of scientific information based on the quality of service and on profit guarantee. In his work, the economic model accounted for CPU time, bandwidth, network, and disk space. Mihailescu [11] presented a dynamic pricing or an auction model for allocating resources in large distributed systems. Woitaszek and Tufo [21] analyzed an economic charging model in the perspective of a supercomputing service provider. They examined IBM Blue Gene/L system cost and applied a charging model that was used in Amazon EC2 [2] to calculate the break-even point. Results showed that pricing was not competitive with the commodity system in Amazon EC2. Xinhui [22] considered not only Cloud TCO but utilization cost as well. The utilization cost was the cost calculated from using part of the resources in order to evaluate efficiency of the Cloud. The utilization cost was calculated according to the number of VMs. Walker [20] presented the real cost of a CPU hour for three cases: purchase, lease and purchase-upgrade. He calculated the Net Present Value and the Net Present Capacity of each case. Niyato [13] proposed economic models to determine an optimal strategy between private Cloud and Cloud service providers for a monopoly market, Nash equilibrium for competitive market, and bargaining for cooperative market.

In this study, we develop an economic model, based on market share and Total Cost of Ownership, for a cloud service provider in order to maximize profit. This study is different from other studies in the literature in that it considers costs due to system failure and repair, to maintain a certain level of availability, and to penalty cost for failing to meet the SLA.

III. REVENUE AND COST ANALYSIS

Price influences the behaviour of customers [10]. If Cloud providers can set the right price, they can obtain higher revenue, and hence more profit. A typical pricing in Cloud services is directly related to the VM hours that a Cloud system provides. If providers can define the demand function of their company or market-share function, they can define the right price in order to have higher revenues. Revenue over a certain time period can be computed as

$$\text{Revenue} = \text{Price} * \text{Quantity} , \quad (1)$$

where Quantity is the number of VM hours that run on the Cloud system and Price is that per VM hour. It is seen from Equation (1) that total revenue is determined by the demand function in Figure 1. A Cloud provider can choose either price or quantity. From the demand function shown in Figure 1, if the Cloud provider sets a price, he or she will know the number of VM hours. On the other hand, if a provider wants to sell a certain quantity (number of VM hours), the Cloud provider has to set price according to the demand function. In this work, we consider three services with different availability. We will utilize three demand functions, based on each service, with a different price.

A logit model is commonly used for representing a demand function [14] or customer behaviour [6], as shown in Figure 1. From the logit demand function in Figure 1, it is seen that when quantity is high, the price is low. The logit model is the model of choice used to model customer behaviour in the market. The logit model is given as

$$D(p) = C * \exp\{-(a + bp)\} / (1 + \exp\{-(a + bp)\}) \quad (2)$$

where p is price, a , b , C are parameters with $C > 0$, $b > 0$. The parameter a can be either less than or larger than 0. Here, C refers to the maximum quantity of the market-share of that firm or its market size. The market price is approximately equal to $-a/b$, [14]. From Equation (2), it is seen that the Total Revenue is

$$\text{Total Revenue} = p * D(p) \quad (3)$$

where $D(p)$ is the number of VM hours.

Cloud TCO is defined in [22] as the cost spent for owning the Cloud system over its life span. It includes the cost of acquiring the Cloud system and operating it. The cost will be classified into eight categories according to [22]: server, network, software, power, cooling, facilities, real estate, and support and maintenance. In the model that we present here, quantity is considered to be the number of servers. A server can run a certain amount of VMs, called VM density [22].

This study is unique in that it considers costs due to system failure and repair, to maintain a certain level of availability, and to penalty cost for failing to meet the SLA. In what follows, we will describe all the costs considered in our model. The Cloud has usually homogenous servers tied into racks. The total cost of servers includes the original service price

plus the upgrade cost to avoid performance degradation in a competitive market. *Server cost* can be computed as

$$\text{Scost} = \text{Nserv} * \text{ServC} + \text{Upgradecost} * \text{year}, \quad (4)$$

where Nserv is the number of servers, ServC is a unit server cost, Upgradecost is upgrade cost per year. In our server cost,

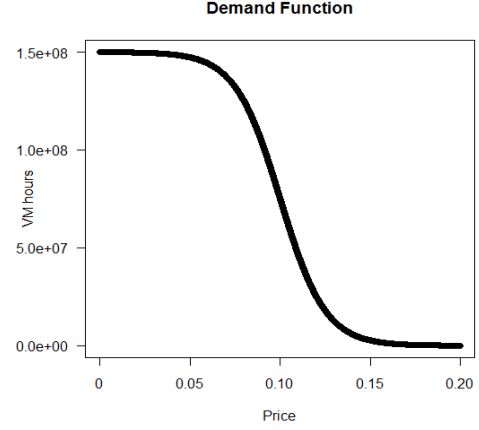


Figure 1. Logit demand function

we consider the upgrade cost (Upgradecost) to be of importance. We define the upgrade cost as follows:

$$\text{Upgradecost} = \text{Nserv} * \text{ServC} * \text{percent}, \quad (5)$$

where percent is 0-100% of server cost.

The *network cost* is mainly the switch cost. The number of switches is based on the number of ports that are needed from the servers. A server can have more than one network interface card (NIC). An NIC may have more than one port. Therefore, *network cost* is defined as

$$\text{Ncost} = \text{SwC} * \text{Nswitch} \quad (6)$$

where SwC is unit switch cost and Nswitch is the number of switches calculated as

$$\text{Nswitch} = \text{Nport} * \text{Nnic} * \text{Nserv} / \text{NSWport}, \quad (7)$$

where Nport is the number of ports per NIC, Nnic is the number of NIC per server, NSWport is the number of ports in a switch.

The *software cost* is mainly the cost for the license. There are Operating systems costs for VMs and software costs for managing VMs. The operating system cost is applied per VM. Managing software is licensed by the number of processors in the system. Therefore, the software cost is

$$\text{SoC} = \text{Nvm} * \text{Oprice} + \text{Npps} * \text{SUprice} * \text{Nserv} \quad (8)$$

where N_{vm} is the number of VMs, O_{price} is the license price per VM, N_{pps} is the number of processors per server, and SU_{price} is software unit price per processor used for management of VMs.

The *power cost* is considered per server. In our model, the power cost is calculated as

$$PowC = N_{serv} * Powpr * v, \quad (9)$$

where $Powpr$ is the power consumption required per server and v is the electric utility cost (kWh).

The *cooling cost* is the power consumption used for cooling the system. In our model, cooling cost is calculated by

$$CoolC = N_{serv} * Powpr * Powpw * v, \quad (10)$$

where $Powpw$ is the power consumption of cooling 1W of heat from the equipment.

Facilities cost is the cost paid for equipment such as KVM switch and cables. They are attached to racks. The facilities cost is calculated by

$$FacC = C_{fac} * N_{rack}, \quad (11)$$

where C_{fac} is cost of facilities per rack and N_{rack} is the number of racks in the system.

Real estate cost is the cost spent for leasing a room required for hosting the servers. Real Estate cost (REC) is computed by

$$REC = C_{sq} * S_{qr}, \quad (12)$$

Here, C_{sq} is cost per square foot per year and S_{qr} is the number of square feet needed by the cloud system. S_{qr} is calculated as

$$S_{qr} = N_{rack} * S_{qRack}, \quad (13)$$

where S_{qRack} is the space in square feet needed by a rack.

The *support and maintenance cost* comes from the salary of the IT staff, performance maintenance, system configuration, and training. Support and maintenance cost is given by

$$SupC = \text{Salary} + \text{training} + \text{CSLA} + \text{RC} + \text{AVC} \quad (14)$$

where Salary is the wage of all IT staff members.

$$\text{Salary} = \text{wage} * N_{rack} / \text{rpIT}, \quad (15)$$

where wage is the salary per IT staff and rpIT is the number of racks that an IT staff handles.

training is the cost spent for training the IT staff, CSLA is the cost spent for penalty due to Service Level Agreement (SLA), repair cost (RC) is cost due to failures in the Cloud, and AVC is the cost of preparing the system for a certain level of availability.

To approximate the *repairing cost*, we assume that the server time to failure (x) follows a gamma distribution

($\Gamma(k_i, s_i)$) and failures are independent [8]. The gamma distribution fits well the TTF of HPC systems, [7], [17], and has a useful property in that the sum of independent gamma random variables is also a gamma random variable. Hence, the repair cost (RC) can be calculated by

$$RC = \sum_{i=1}^{N_{serv}} NF_i * AC, \quad (16)$$

where AC is average repair cost per failure and NF is the expected number of failures per server. The expected number of failures can be obtained as $E[\frac{T}{x}]$, where T is the lifespan of the Cloud and x is the time to failure of a server. $E[\frac{T}{x}]$ can be approximated using the Taylor series expansion at $a=\mu$.

$$E[\frac{T}{x}] = E[\frac{T}{\mu} - \frac{T}{\mu^2} (x - \mu) + \frac{T}{\mu^3} (x - \mu)^2] = \frac{T}{\mu} + \frac{T * \text{Var}(x)}{\mu^3} \quad (17)$$

Here, μ is mean time to failure and can be obtained from the gamma distribution.

Cloud customers may require different system availabilities. For example, mission critical applications may need 99.999% availability, while other not so critical applications may require a lower availability level. Cloud service providers have to engineer the process of repair time in order to meet a given level of availability. The costs that are usually encountered are the cost for reducing time-to-repair, the cost for redundant equipment such as servers and switches, and the cost for hiring and training. Scott [18] has shown that a higher cost is associated with achieving higher levels of availability. He also showed that cost rises exponentially with an increase in the availability level. In this work, we follow Scott [18] in assuming an exponential relationship between availability and cost as shown in Figure 2.

The *availability cost* for each availability level is calculated as follows:

$$\text{AVC} = \exp\{\text{availability level}\} * \text{HAC}, \quad (18)$$

where availability level is the level of availability such as 99.999, 99.99 or 99.95, HAC is the highest availability cost for engineering the Cloud system to have a 99.999% availability. To achieve 99.999% availability, one has to have redundant servers and sufficient well-trained IT staffs. Therefore, HAC is a cost to the system, which considers redundant server cost (RScost), availability software cost (ASoC), additional switch cost (NCost) for redundant server, additional IT Staff (salary) cost, training cost, additional facility cost, and additional space cost (AREC) (for redundant servers). HAC is computed by

$$\text{HAC} = N_{serv} * \text{RScost} + \text{ASoC} + \text{NCos} + \text{salary} + \text{training} + \text{FacC} + \text{SpaceC} + \text{AREC} + \text{PowC} + \text{Cool}, \quad (19)$$

Availability software is the module that manages server redundancy for maintaining a certain level of availability in a Cloud system. Availability software cost (ASoC) is calculated by

$$\text{ASoC} = \text{Npps} * \text{ASUprice} * \text{Nserv}, \quad (20)$$

where Npps is the number of processors per server, and ASUprice is availability Software price per processor.

For *penalty cost*, SLA is used for guaranteeing the satisfaction of a Cloud customer. Quality (throughput and response time) [12] and reliability [2], [15] of Cloud services are of main concern. Cloud service providers such as Amazon EC2 and Rackspace [15] guarantee certain availability for customers. If availability is less than the point they claim, customers receive service credit back. Therefore, in this analysis, we will approximate the cost coming from SLA due to refunds to customers as a result of failure to meet the guaranteed level of availability.

Availability is uptime divided by the total operation time. Given A% availability of a VM, one can calculate downtime (t) per year. If downtime exceeds t hours per year, then the Cloud provider would encounter cost due to customer refunds. Therefore, one would like to know the probability (P_D) that the total downtime in a year is more than t hours (that is the probability of encountering costs due to refunds). One can calculate this probability for a VM as follows:

$$P_D = \sum_{r=1}^{\infty} \{1 - P[T_r \leq t]\} * P[N(y) = r] \quad (21)$$

where r is the number of failures, T_r is the sum of repair times ($t_1 + \dots + t_r$), and $N(y)$ [4] is the number of failures in a year interval. Assume that repair time follows a gamma distribution $\Gamma(\alpha, \beta)$ [16] and is independent of other repair times. T_r is the sum of r independent gamma random variables, therefore, T_r has a gamma distribution, $\Gamma(r*\alpha, \beta)$. Hence, one can compute $P[T_r \leq t]$ from $\Gamma(r*\alpha, \beta)$. We assume that time to failure (TTF) follows a gamma distribution [17]. Hence,

$$P[N(y) = r] = e^{-t/s} \sum_{i=ra}^{(r+1)(a-1)} \frac{t^i}{s^i i!} \quad (22)$$

In practice, the gamma parameters α, β can be estimated from data on repair time. Also, the parameters a and s can be estimated from data on time to failure of a server.

The expected *penalty cost* due to customer refunds (CSLA in Equation (14) is calculated by

$$\text{CSLA} = \sum_{i=1}^{\text{Nserv} * \text{Nvm}} P_{D_i} * \text{SC}, \quad (23)$$

where SC is the service credit that is refunded to customers when availability of the VM is less than A%, and Nvm is the number of VMs that run on a physical machine.

Total cost is the summation of eight costs over the lifetime of the Cloud. This can be expressed as

$$\text{Total cost} = \text{Scost} + \text{Ncost} + \text{SoC} + \text{PowC} + \text{CoolC} + \text{FacC} + \text{REC} + \text{SupC} \quad (24)$$

IV. ANALYSIS AND RESULT

In this section, we demonstrate with an example how to maximize profit based on demand and cost analysis, by choosing the right price and rightsize of a Cloud. After revenue and cost are analyzed, profit can be calculated from the following formula

$$\text{Profit} = \text{Revenue} - \text{Total Cost}, \quad (25)$$

where Revenue is from Equation (3) and Total Cost from Equation (24). Revenue as well as cost were calculated for each of the availability levels in Figure 2. In this Figure, the number of servers was determined by the level of demand in VM hours, (D(p)) for a given price, p. The number of servers was then used to determine the total cost from Equation (24). To maximize profit, one can readily determine the number of servers that maximizes the difference between total revenue [p(D(p))] and total cost over the lifespan of a certain number of years.

We consider revenue and cost over five years. Also, we consider three types of services that have different Quality of Service (QoS), which guarantee 99.999%, 99.99% and 99.95% availability. We assume that the maximum market share (C) of the Cloud provider is 1.0×10^8 , 1.25×10^8 , and 1.5×10^8 VM hours per year and the market price is \$0.20, \$0.15, and \$0.10 per VM hour [20] for each of the availability levels, respectively. We consider the server capacity as Quantity in Equation (1). For the Cloud cost analysis, we assume a 42 1U compute blade racks. A dual-processor compute blade server costs \$2000 [20]. The upgrade cost is equivalent to the original server cost [20]. Also, the redundant server cost was 1.5 times the original server cost and the number of VM per server was 12. We estimate that a server requires 0.6 kW to operate, and 0.6 W for cooling. We assume that it requires 1W to cool 1W of heat from the equipment (Powpw). A rack needs around 20 square feet (SqRack), which costs approximately \$400 per square foot (Csq) [22]. A server has 2 NICs and each NIC has 2 ports. A 48-port Gigabit Router costs around \$3000. For the software cost, we assume that the Cloud uses Linux-based operating systems in VMs. The managing software cost and availability software per processor is \$100. We assume 4 racks per IT staff. The salary for an IT staff is \$5000 a month. Training IT staff per year costs \$250. Repairing cost is calculated by Cost per time (\$200) [3], [19] multiplied by the number of failures. We consider that mean time-to-failure (MTTF) is 2184 hours with variance equals to 1657 [7]. We assume no downtime for 99.999% availability service because of redundancy. Mean-time-to-repair is 7.5 and 30 minutes for 99.99% and 99.95%, respectively.

Results are presented in Figure 2 for each service level agreement. Figure 2 shows revenue and cost for each availability level. Cost is a linear function because it is based only on the number of servers. When the number of servers increases, the cost increases. To find the number of servers that maximizes the profit, one chooses the number that maximizes the difference between revenue and cost, Figure 2. For 99.999% availability service, the maximum profit is

attained for 798 servers with a price of \$0.182 per hour. For 99.99% availability service, 840 servers give the maximum profit with a price of \$0.143 per hour. For 99.95% availability service, the profit is maximized with 1004 servers and a price of \$0.8919 per hour.

Figures 3 (a) and (b) present price per VM hour and number of servers that maximize profit for a given server cost and availability. Figure 3 (c) shows the maximum profit (total revenue – total cost) for a given server cost for 99.95%, 99.99% and 99.999% availability. As seen from these graphs, as cost per server increases, the number of servers that maximizes profit decreases, price per VM hour that maximize profit increases, and maximum profit decreases. This analysis helps a Cloud service provider determine price per VM hour and number of servers that would maximize profit, given a certain cost per server. For instance, for 99.95% availability shown in Figure 3 (a), (b), and (c), if server cost increases from 1,000 to 2,000 where the slope of price per VM hour and the number of servers are flat, the Cloud service provider should keep the same price per VM hour as well as the same number of servers in order to maximize profit. Notice, however, that maximum profit for \$2000 cost is less than that for \$1000 cost

V. CONCLUSION

Finding the right size and the right price is crucial for a Cloud service provider. Rightsizing and right pricing of the Cloud are based on revenue and cost analysis. Pricing is closely related to revenue. If the providers define the price per VM hour higher than the market price, fewer customers are willing to pay. On the other hand, if the price is very low, the cost will exceed the revenue. How to price a VM hour is important. The Total Cost of Ownership (TCO) is a tool used to analyse the cost of owning the Cloud. In this paper, we proposed an economic model for Cloud service providers that can be used to maximize profit based on choosing the rightsizing in the Cloud data center for three different qualities of service. We demonstrate, through an example, how to obtain maximum profit based on right pricing and rightsizing. This work can be extended by considering revenue from charging for storage and data transfer. In addition, customer satisfaction can be explored as part of the future model.

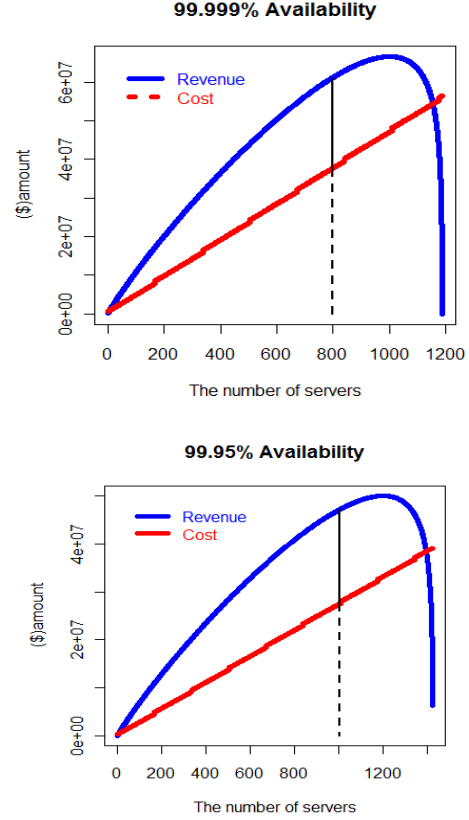
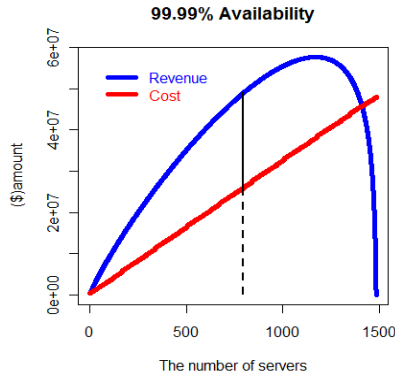
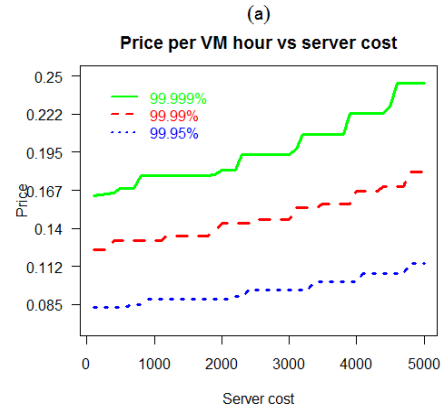


Figure 2. Total revenue [pD(p), Equation (3)] versus total cost (Equation (24)) with 99.999, 99.99, and 99.95% availability level.



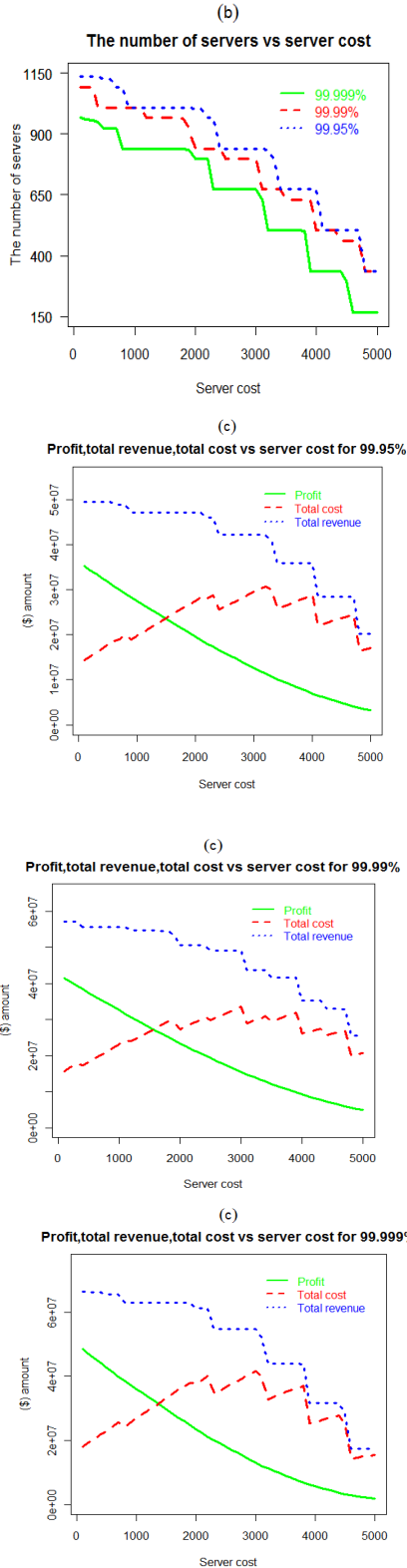


Figure 3 a,b,c. (a) Price per VM hour that maximizes profit for a given server cost for each availability. (b) the number of servers that maximizes profit for a given server cost for each availability. (c) Maximum profit (revenue - total cost) for a given server cost for 99.95%, 99.99%, 99.999% availability.

VI. REFERENCES

- [1] Abdelkader, K., Broeckhove, J., Vanmechelen, K., "Commodity Resource Pricing in Dynamic Computational Grids" IEEE/ACS, (2008), 422-429.
- [2] Amazon EC2. <http://aws.amazon.com/ec2/> (last accessed on January 25, (2012))
- [3] Barroso, L.A., Holzle, U., "The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines", In Synthesis Lectures on Computer Architecture, (2009).
- [4] Chiang, C.L., "An Introduction to Stochastic Processes and Their Applications", Robert E. Krieger Publishing Co., INC., New York, (1980).
- [5] Dash, D., Kantere, V., and Ailamaki, A., "An Economic Model for Self-Tuned Cloud Caching", In Proceedings of the 2009 IEEE International Conference on Data Engineering (ICDE '09). IEEE Computer Society, Washington, DC, USA, (2009), 1687-1693.
- [6] Dube, P., Hayel, Y., and Wynter, L., "Yield management for IT resources on demand: analysis and validation of a new paradigm for managing computing centers", Journal of Revenue and Pricing management, (2005), 4(1), 24-38.
- [7] Gottumukkala, N. R., "Failure Analysis and Reliability-Aware Resource Allocation for Parallel Applications in High Performance Computing Systems", Ph.D. Dissertation. Louisiana Technical Univ., Ruston, LA, USA, (2008), AAI3298937.
- [8] Gottumukkala, N. R., Nassar, R., Paun, M., Leangsuksun, C. B., Scott, S. L., "Reliability of a System of k Nodes for High Performance Computing Applications", IEEE Transactions on Reliability, (2010), 59(1), 162 - 169.
- [9] Hogan, J. and Nagel, T., "The Strategy and Tactics of Pricing: A Guide to Growing More Profitably", 4th edition, Pearson Prentice Hall, (2006).
- [10] Landsburg, S. E., "Pricing Theory and its Application", Thomson, 7th edition, (2008).
- [11] Mihailescu, M., Teo, Y. M., "On Economic and Computational-efficient Resource Pricing in Large Distributed Systems", Cluster, Cloud and Grid Computing (CCGrid), 10th IEEE/ACM International Conference, (2010), 823-843.
- [12] Moon, H. J., Chi, Y., Hacigümüs, H., "SLA-Aware Profit Optimization in Cloud Services via Resource Scheduling", IEEE SERVICES, (2010), 152-153.
- [13] Niyato, D., Chaisiri, S., Bu-Sung Lee, "Economic Analysis of Resource Market in Cloud Computing Environment", Services Computing Conference, APSCC 2009. IEEE Asia-Pacific, (2009), 156-162.
- [14] Phillips, R., "Pricing and Revenue Optimization", Stanford University Press, (2005).
- [15] Rackspace: <http://www.rackspace.com> (last accessed on January 25, (2012))
- [16] Sarala Arunagiri, John T. Daly, and Patricia J. Teller, "Propitious Checkpoint Intervals to Improve System Performance", Technical Report UTEP-CS-09-09, (2009).
- [17] Schroeder, B., Gibson, G.A., "A large-scale study of failures in high-performance computing systems", IEEE Dependable and Secure Computing, (2010), 7(4), 337-351.
- [18] Scott, D., "The High Cost of Achieving Higher Levels of Availability", Strategic Planning, SPA-13-9852. Research Note, (2001).
- [19] Vishwanath, K. V., Nagappan, N., "Characterizing cloud computing hardware reliability", In Proceedings of the 1st ACM symposium on Cloud computing (SoCC '10). ACM, New York, NY, USA, (2010), 193-204.
- [20] Walker, E., "The Real Cost of a CPU Hour", IEEE Computer Journal, (2009), 42(4), 35-41.
- [21] Woitaszek, M., Tufo, H.M., "Developing a Cloud Computing Charging Model for High-Performance Computing Resources", Computer and Information Technology (CIT), IEEE 10th International Conference, (2010), 210-217.
- Xinhui, L., Ying, L., Tiancheng, L., Jie, Q., Fengchun, W., "The Method and Tool of Cost Analysis for Cloud Computing," cloud, IEEE International Conference on Cloud Computing, (2009), 93-100.