2015 International Conference on Cloud and Autonomic Computing

Optimal Pricing and Capacity Planning of a New Economy Cloud Computing Service Class

Jie Xu
Department of System Engineering & Operations Research
George Mason University
4400 University Drive, MS 4A6
Fairfax VA 22030, USA
Email: jxu13@gmu.edu

Chenbo Zhu
College of Economics and Management
Zhejiang University of Technology
Hangzhou 310023, China
Email: chenbozhu@zjut.edu.cn

***Abstract*—Resource under-utilization in cloud computing systems is widespread due to workload fluctuations and drives up the cost of cloud computing service. Offering service using slack resources in an opportunistic way improves the utilization of resources and the economics of cloud service providers. But Opportunistic service class comes with virtually no service level objectives (SLO) and thus is of limited use. In a recent study, a new Economy class was introduced to provide long-term SLOs using reclaimed cloud computing resources. Analysis based on the workload collected on six production cloud computing clusters at Google demonstrated the potential of the Economy class. This paper presents an analytic study on the optimal pricing and capacity planning of this new Economy class. We show that depending on the terms of the service level agreements and the characteristics of the cloud computing workloads, a cloud service provider may either choose a penalty averse or penalty preference strategy when allocating reclaimed computing resources to the Economy class cloud computing service. We also derive conditions under which the new Economy class will be profitable.**

I. Introduction

Infrastructure-as-a-Service (IaaS) cloud provides users with affordable and elastic computing service. Essential to this affordability and elasticity are virtualization technology and statistical multiplexing. Virtualization technology enables the cloud service operators to provision virtual machines (VMs) instead of physical servers to host different applications. Each VM is allocated a certain amount of resources and multiple VMs can be placed on the same physical server. Statistical multiplexing exploits the reduction of the variability of aggregated workload fluctuations and allows the cloud operator to provision physical resources that are less than users' total requests for resources. Workload statistics collected from six Google production cloud computing clusters from December 2012 to November 2013 help make this point [1].

The CPU ceiling utilization, which is defined as the ratio of the user's actually requested resource limits to the maximum amount of resources that the cloud provider grants, varies significantly at user level, with many users having a ceiling utilization between 30% and 63%. But for 38% users, the CPU ceiling utilization is no more than 1%, while the CPU ceiling utilization exceeds 99% for 15% of the users. In contrast, the CPU ceiling utilization at the cluster level, which is defined as the sum of the total requested CPU resources over the sum of all ceilings, was much more stable, varying between 55% and 75% most of the time. Only less than 1% workload measurements show a cluster level CPU ceiling utilization lower than 1%, or higher than 81% [1]. Because of this decrease in workload variability, it is much easier for cloud operators to use statistical multiplexing and use a server to host VMs with total allocated capacity exceeding the physical capacity of the server [2], [3].

While statistical multiplexing allows the cloud operator to increase resource utilization and thus its profitability, it also leads to service unavailability when all or a large proportion of users' workloads peak at the same time. When this happens, some VMs would not be able to access resources allocated to them. The more aggressively a cloud operator applies statistical multiplexing, the more frequently service unavailability would occur. To provide a long-term, e.g., monthly, availability at a level of 99.95% or 99%, as specified in Amazon EC2 service level agreement (SLA) [4], cloud operators have to maintain enough physical capacity and limit the extent of statistical multiplexing and its economic benefits. Consequently, data center resource utilization rates are typically quite low. For example, in the Google cloud computing clusters studied in [1], total resource slack accounts for about 57% of the cloud computing cluster capacity. This represents a significant waste of resources and hurts the profitability of cloud operators.

One way to increase resource utilization in data centers is to offer unused resources to users at a much lower price but in an opportunistic way [5]. Such opportunistic cloud service would virtually have no SLO as the resources could be preempted at any moment to be used for other VMs with SLOs. One such example of opportunistic cloud service is Amazon EC2 Spot Instances. Amazon EC2 Spot Instances ask users to specify the maximum hourly price that the user is willing to pay for running VMs [6]. Amazon EC2 specifies a current Spot Price, which changes dynamically to reflect supply and demand. Whenever the user's bid price exceeds the current Spot Price, the user's VM instances would run; whenever the Spot Price exceeds the user's maximum rate, the VMs would be shut down. Spot instances may provide the user with tremendous cost savings, which is on average 86% on Amazon EC2 recently [6]. However, there is clearly no Service Level Objectives (SLO) for such opportunistic cloud service. This limits this type of cloud service to only computing tasks that last a very short amount of time or are very flexible in response time and can tolerate frequent interruptions.

/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCAC.2015.8

In [1], a new type of cloud service class, referred to as the Economy class, was introduced to provide long-term SLOs using reclaimed cloud computing resources. The Economy cloud computing service class, which we will simply refer to as the Economy class for convenience hereafter, uses reclaimed resources to provide long-term SLOs that are only slightly weaker than those of the On-demand/Reserved cloud computing service classes, e.g., 98.9% vs. 99.95% for a six-month period. But the price of the Economy class will be lower than that of the On-demand/Reserved class. When there are SLO violations, cloud service providers pay a financial penalty to users as stipulated in the SLA of the Economy class. The rest of the reclaimed resources are used to offer the Opportunistic class service at the lowest price with virtually no SLO.

By constructing prediction confidence intervals for the total resource slack in a future six-month time period, the authors in [1] studied via a sensitivity analysis the trade-offs between aggressively and conservatively using reclaimed cloud computing resources for the new Economy class. It was shown using real world Google production cloud computing cluster data that it was possible to offer the Economy class with reasonably high six-month SLOs (e.g., 98.9%). This makes the Economy class service attractive for applications that can tolerate slightly worse availability SLOs, such as web indexing and video transcoding [1]. The authors experimented with different prices for the Economy class and illustrated that using reclaimed resources to provide the new Economy class cloud service can significantly increase data center resource utilization and the profitability of the cloud operator [1].

In this paper, we build on the work of [1] and present a mathematical and economic analysis of the new Economy cloud computing service class. Our goal is to answer three questions that are critical to the success of the Economy class as a new type of cloud computing service: capacity planning, pricing, and market profitability.

The rest of the paper is organized as follows. In Section II, we review related work on resource management in cloud computing systems. We then describe the problem setting in Section III and explicitly list the notations that we will later use in Section IV to develop a mathematical model. We present our model and solutions to capacity planning and pricing in Section IV, followed by results on the profitability of the Economy class in Section V. We conclude the paper in Section VI.

II. Related work

Resource management in data centers have been extensively studied with different objectives. Many studies try to improve the energy efficiency with various virtualization techniques. Methods such as VM consolidation [7]–[12], VM migration [13], [14], and dynamic VM provisioning [15], [16] aim to use as few servers as possible to host VMs. Such a strategy is supported by past studies that show the fixed part of energy consumption once the server is on dominates the part of the energy consumption that varies according to the workloads [17], [18]. It has been shown that when properly managed, moving severs offline or putting them to a sleep state can reduce energy consumption cost without sacrificing SLOs [19]–[21]. There are also studies that examine other cost factors in data center operations [22], [23]. All these earlier works contribute to improving the profitability of a cloud operator by focusing on reducing data center operations cost. Compared to these studies, this paper focuses on analyzing how the additional revenue generated by the new Economy class can help improve the profitability of IaaS cloud service.

Amazon EC2 Spot Instances provides an industrial product that makes use of reclaimed cloud computing resources to improve resource utilization and revenue [6]. Because Spot instances are shut down when the user's bid price is lower than the current Spot Price, which is specified by Amazon EC2 and beyond the control of the user, there is basically no SLO for Spot instances. To gain better service availability, users can increase bid prices but that diminishes the cost savings from using Spot Instances and still provides no SLOs. In the literature, using reclaimed resources in an opportunistic way was described in [5], where VMs running on reclaimed resources can be preempted and thus no SLO is offered essentially. In [1], the Economy class with long-term SLOs using reclaimed cloud computing resources was introduced and this work builds on [1]. We provide a mathematical model to address key design questions related to the Economy class.

Using reclaimed resources requires one to predict the fluctuations of workloads in the data centers. In [24]–[26], time series prediction models are proposed to generate point forecast values for data center workloads. The forecast time windows are typically quite small, e.g., a few hours or at most days. In [1], since the Economy class is envisioned to provide long-term, e.g., six-month SLOs, time series techniques such as ARIMA and ETS were employed to forecast long-term slack resource availability. To help study the tradeoffs between aggressive and conservative capacity planning strategies for the Economy class, prediction intervals were constructed using techniques in [27]. There have also been studies focusing on workload predictions for specific types of tasks. In [28]–[30], forecasting techniques for video streaming demands were proposed and evaluated.

As we shall see in Section IV, our model requires knowledge of workload distributions. Therefore, point forecasts are not useful in our context. However, the prediction distributions described in [27], combined with the analysis reported in [1], provide the tools to generate the required predictions for our model. But such discussion is beyond the scope of this paper, which focuses on the development of the mathematical model and solution to the pricing and capacity planning for the Economy class.

Also related is the work on the analysis of the workloads and resource consumption patterns of different tasks. Studies using Google cloud cluster data traces provide detailed information on real world workloads and tasks. In [31], [32], the authors classified computing tasks based on the durations and resource consumption patterns. In [33], the heterogeneous and dynamic nature of cloud resource requests was revealed using Google cluster data traces, showing substantial variations in requested resource configurations, e.g., different ratios of processors, memory, and storage, as well as variations in physical server configurations. The heterogeneity in resource requests and physical server configurations leads to both challenges and opportunities in research on using reclaimed cloud computing

resources to provide long-term SLOs via the Economy class. However, we will pursue such a study in the future and focus on a pre-defined resource bundle in the current work, as did in [1].

III. Problem descriptions and notations

We consider a cloud computing resources management problem in an IaaS cloud service provider. In an IaaS cloud, users send requests for VMs to cloud service providers to meet their computing demands. Cloud service providers offer different configurations designed to satisfy the different demands of users. For instance, Amazon EC2 provides General Purpose (T2, M3), Compute Optimized (C3, C4), Memory Optimized (R3), GPU (G2), and Storage Optimized (I2, D2) configurations [34]. In this paper, we make the simplifying assumption that there is only one VM configuration offered, as did in [1].

Users can request up to a maximum number of VM instances, which is referred to as the *ceiling*. For example, Amazon EC2 has a default ceiling of 20 instances for On-Demand and Reserved Instances and 5 instances for Spot Instances. Users can send requests to increase their ceilings. When a user makes a VM request and the total requested amount of resources is lower than the users's ceiling, the IaaS cloud service provider accepts this request, if there is enough capacity remaining, and allocates resources to host the VM(s). A cloud operator may also reserve capacity in anticipation of new VM requests from a user within the users' ceiling.

An IaaS cloud service provider often offers several service classes to users with different SLOs and prices to satisfy the diverse demand for service quality and cost among users. Take Amazon EC2 for instance. On-Demand Instances and Reserved Instances come with high SLOs. Amazon promises to make EC2 available with a monthly uptime percentage of at least 99.95% [4]. If there are SLO violations, Amazon will apply a service credit to affected users' monthly bills, i.e., when the monthly uptime percentage is below 99.0%, there will be a 30% service credit. The main difference between EC2 On-Demand Instances and Reserved Instances is the Reserved Instances require long-term commitments but give users a significant discount (up to 75%) compared to On-Demand Instance pricing [34]. In this paper, we consider both Reserved service class and On-Demand service class as one service class that provides high quality SLOs, and will simply refer to it as the Reserved service class. The IaaS service may preempt resources allocated to other service classes when necessary to avoid SLO violations for Reserved service class users.

The IaaS cloud operator's resources will not be fully utilized by the Reserved class; otherwise, there must be frequent SLO violations. In [1], unused resources are classified into three categories:

- Reservation slack: Reservation slack includes capacity that has not been reserved and is thus free to be used to host VM requests from other service classes.
- Allocation slack: Allocation slack refers to capacity that has been reserved, i.e., in anticipation of future VM requests from the user, but has not been allocated to specific running VMs. Allocation slack thus represents capacity that can be easily reclaimed and used for other service classes, but is likely to be preempted and thus becomes unavailable to the Economy or Opportunistic class service.
- Usage slack: Usage slack is a result of the stochastic fluctuations in resource consumptions among running VMs. For example, the users' VM may be idling and thus does not consume any physical resource in the intermission between two computing tasks. This type of slack is the most difficult to reclaim because not only it is more likely to be requested again by the running VM, but it requires special techniques to make such resources available to other VMs, e.g., balloon drivers [35].

Our focus is on how the IaaS cloud service provider can make use of these slack resources to increase resource utilization and profitability. We do not study the technology required to reclaim different types of resource slacks and thus treat them as the same for our purpose. We assume that the IaaS cloud service provider currently offers an opportunistic service class using reclaimed resources but there is no availability SLO for the opportunistic service class. The lack of SLOs makes this service class not suitable for many applications.

We want to study the impact of adding a new Economy service class as explained in [1]. The new Economy class also uses reclaimed resources but come with strong long-term SLOs, making it an attractive choice to users that can accept long-term SLOs slightly weaker than those of the Reserved class, e.g., 98.9% vs. 99.95% but want a price lower than that of the Reserved service class. Our analysis attempts to address three critical questions related to the success of the new Economy service class:

- What is the optimal proportion of reclaimed resources to allocate to the new Economy class?
- What is the optimal price for the Economy class?
- Under what conditions will offering the Economy class increase the profit of the IaaS cloud service provider?

We assume that the only SLO specified in the SLA is VM availability. For example, Amazon EC2 SLA defines "Unavailability" as "when all of your running instances have no external connectivity" and SLO violation is defined as a monthly unavailability less than 99.95% for On-Demand and Reserved instances.

Similar to [1], we make the simplifying assumption that there is enough demand to consume all resources. We further assume that none of the SLO violation, if there is any SLO violation at all, for the Reserved class is caused by the Economy or Opportunistic classes. This is a reasonable assumption as the Iaas cloud service provider can avoid such SLO violations by preempting resources from the Economy and Opportunistic class instances. Because of this assumption, it is not necessary to model the penalty for Reserved class SLO violations and we will assume that there is no such penalty incurred for notational simplicity. Before we proceed with our mathematical model and analysis, we first list the notations used in the rest of the paper in the following. We begin with the list of parameters that are considered known and fixed in the following. We

follow the same structure of SLA as in [1] for the Economy class, which specifies that if there is an SLO violation in a billing period, the IaaS cloud service provider incurs a penalty as a percentage of the user's service bill in that billing period.

- p_o: the price per unit of allocated resource for the Opportunistic class;
- A_r: the availability SLO of the Reserved class, e.g., 99.95%;
- A_e: the availability SLO of the Economic class, e.g., 98.9%;
- X: the penalty paid to the Economy class user if the availability SLO is violated, as a discount applied to the Economy class user's service bill;
- V_r: marginal valuation of the Reserved class service with the availability SLO A_r;
- V_e: marginal valuation of the Economy class service with the availability SLO A_e;
- $u \in [0, 1]$: the resource usage rate for both the Reserved class and the Economic class during the time period under consideration;

Below is then a list of decision variables that examines the pricing of the Reserved class and the Economy class, and the capacity allocation decision for the Economy class:

- p_r: the price per unit of allocated resource for the Reserved class;
- p_e: the price per unit of allocated resource for the Economic class;
- α: the proportion of the average amount of reclaimed resources allocated to the Economy class.

The parameters A_e and X are SLA parameters as determined by the IaaS cloud service provider. The parameters V_r and V_e in practice will be estimated through marketing research conducted on the targeted IaaS cloud service user segments. Following standard price in market segmentation analysis [36], we assume the customer segments are homogeneous and thus V_r and V_e are constant for all customers belonging to the targeted segments. The valuations of the Reserved and Economy cloud service classes are then given by $V_r A_r$ and $V_e A_e$ [36]. The price of the Opportunistic class p_o and the availability SLO for the Reserved class A_r are also assumed to be given prior to the analysis, i.e., according to the cloud service provider's previous marketing research.

We denote the mean of u by μ_u. We assume that u has a symmetric probability density function $f(\cdot)$ and a cumulative distribution function $F(\cdot)$. Notice that the widely used normal distribution in time series model [27] has a symmetric probability density function. We further define the mean of resource slack $\bar{u} = 1 - u$ as $\mu_{\bar{u}}$. Notice that $\mu_{\bar{u}}$ is the average amount of reclaimed resources in the data center. In practice, $f(\cdot)$ and $F(\cdot)$ can be obtained from future workload forecast via time series methods [27].

We normalize both the total capacity of the data center and the time duration under consideration to one unit for simplicity. Similar to [1], we will ignore data center operations cost for two reasons. First, it is difficult to accurately calculate a data center operator's cost. Second, there is no evidence to support the conjecture that operations cost would change for different SLOs and thus the operations cost for the three different service classes may be assumed to be about the same. We can thus focus on maximizing the total revenue generated by the different service classes for the Iaas cloud service provider. We now move on to present our models and analysis in the next section.

IV. Models and solutions

We first model the IaaS cloud service provider's revenue when there are only Reserved class and Opportunistic class. The problem to optimize the revenue can be formulated as follows:

$$\begin{array}{ll} \max\limits_{p_r} & \pi_{ro} = p_r + p_o \int_0^1 (1-u) f(u) du, \\ \text{subject to} & V_r A_r - p_r \geq 0, \\ & p_r \geq p_o \geq 0. \end{array} \tag{1}$$

In the objective function in (1), the first term is the revenue from the Reserved class. Recall that we normalize capacity and time duration to one unit. The second term is the average revenue from offering the Opportunistic class using the reclaimed resources, which depends on the probability distribution function of the Reserve class resource usage. The second constraint specifies that the Reserved class is priced higher than the Opportunistic class. The first constraint is known as customer self-selection constraint in the economics and marketing science literature [36]. The meaning of the constraint is the utility of the service to the targeted customer segment, as measured by $V_r A_r$, must be larger than or equal to the perceived value of the service A_r.

It is straightforward to derive that the optimal p_r is the highest price supported by the Reserved class users' valuation of the service, which is shown as follows:

$$p_r^* = V_r A_r. \tag{2}$$

Therefore, the optimal total revenue of the service provider can be calculated as

$$\pi_{ro}^* = V_r A_r + p_o \mu_{\bar{u}}. \tag{3}$$

Notice that resource consumption by the Reserved class is u as we normalize total system capacity to one unit. Recall the mean resource slack is $\mu_{\bar{u}}$ and the Economy class is thus allocated $\alpha\mu_{\bar{u}}$ resources and utilize $\alpha\mu_{\bar{u}}$ of the allocated resource slack. Therefore, $\hat{u} = 1 - (1 + \alpha\mu_{\bar{u}})u$ gives the amount of reclaimed resources left over for the Opportunistic class after offering the Economy class cloud service using reclaimed resources. We now consider the problem of maximizing the revenue of an IaaS cloud service provider offering the Reserved, Economy, and Opportunistic service classes. The optimization problem is formulated as follows:

$$\begin{array}{ll} \max\limits_{\alpha, p_r, p_e} & \pi_{reo} = p_r + p_o \int_0^{\frac{1}{1+\alpha\mu_{\bar{u}}}} \hat{u} f(u) du \\ & \quad + p_e \alpha \mu_{\bar{u}} \left(1 - XI\{\Pr[\hat{u} \geq 0] < A_e\}\right) \\ \text{subject to} & V_r A_r - p_r \geq 0 \\ & V_e A_e - p_e \geq 0 \\ & V_r A_r - p_r \geq V_r A_e - p_e \\ & V_e A_e - p_e \geq V_e A_r - p_r \\ & p_r \geq p_e \geq p_o \geq 0 \end{array} \tag{4}$$

In the new formulation (4), the first term is still the revenue from the Reserved class. The second term is now the revenue from the Opportunistic class using the amount of reclaimed resources left over from the Reserved and Economy class instances usage. The third term represents the revenue from the Economy class. Notice that it includes an indicator function that applies a financial penalty when the IaaS cloud service provider is not able to avoid Economy class SLO violations on average:

$$I\{\Pr[\hat{u} \geq 0] < A_e\} = \begin{cases} 1, & \text{if } \Pr[\hat{u} \geq 0] < A_e; \\ 0, & \text{if } \Pr[\hat{u} \geq 0] \geq A_e. \end{cases}$$

In (4), the last constraint again ensures the right pricing hierarchy, with the Reserved class priced the highest, the Economy class in the middle, and the Opportunistic class the lowest priced. Similar to the constraints in the previous model, the perceived utility of the service designed for the targeted user segment must exceed the price for the service, which is indicated by the first two constraints for the Reserved and Economy classes.

The other two constraints do not exist in the previous model. These constraints are introduced to enforce self selection of services by targeted user segments, as was originally introduced in the literature of customer segmentation [36]. In this approach, it is assumed that the features of a product/service are designed such that only the targeted customer segment finds the product/service attractive. In our context, it means that the availability SLOs A_r and A_e, and the lack of SLO for the Opportunistic class, are distinct enough to target different user segments. However, the targeted user segments also need to perceive their own service as fairly priced.

For example, in these constraints, $V_r A_r - p_r$ measures the net utility that a Reserved class user perceives from using the Reserved service class. When the Reserved class user checks the SLA terms of the Economy price, the user perceives a net utility of $V_r A_e - p_e$. While the Reserved class user would not choose the Economy class because of the user's "hard" requirement of higher availability SLO, the perceived net utility of the Economy class cloud service should not be higher than that of the Reserved class cloud service. Otherwise, Reserved class users would feel the Reserved class is being unfairly priced and thus would lead to a failure of the market segmentation approach and the business model.

Similar to model (1), we can obtain the optimal p_e and p_r, which are shown as follows:

$$p_e^* = V_e A_e, \tag{5}$$

and

$$p_r^* = V_r A_r - (V_r - V_e) A_e. \tag{6}$$

Not surprisingly, p_e^* and p_r^* call for the highest prices that satisfy all the user self-selection constraints and pricing hierarchy to maximize the IaaS service provider's revenue. We can then plug in p_e^* and p_r^* to optimize the objective function only:

$$\begin{aligned} \max_{\alpha} \quad & \pi_{reo} = V_r A_r - (V_r - V_e) A_e + p_o \textstyle\int_0^{\frac{1}{1+\alpha\mu_{\bar{u}}}} \hat{u} f(u) du \\ & + V_e A_e \alpha \mu_{\bar{u}} \left(1 - X I\{\Pr[\hat{u} \geq 0] < A_e\}\right) \end{aligned}$$

Further analysis to obtain the optimal capacity planning strategy α requires differentiating two cases corresponding to how aggressively or conservatively the IaaS service provider allocates reclaimed resources to the Economy class and trades off additional revenue with financial penalty. We separate the discussion into a *penalty averse* strategy and a *penalty preference* strategy.

A. Penalty-averse Strategy

If $\Pr[\hat{u} \geq 0] \geq A_e$, it means that the cloud service provider is able to meet the availability SLO on average and avoid paying for the Economy class SLO violation. This is done by ensuring there is enough spare capacity after allocating $\alpha \times 100\%$ of the average amount of reclaimed resources to the Economy class to cushion resource usage fluctuations. Notice that this is a probabilistic calculation and does not mean that the Economy class availability SLO would never be violated for any user during a billing period. Instead, it means that on average the IaaS cloud service provider would be able to fulfill the SLO.

In this case, the problem can be further simplified and the optimal α can be calculated directly by solving $\Pr[1 - (1 + \alpha\mu_{\bar{u}})u \geq 0] = A_e$, which leads to:

$$\alpha_1^* = \frac{1}{F^{-1}(A_e)\mu_{\bar{u}}} - \frac{1}{\mu_{\bar{u}}}. \tag{7}$$

We refer to such a capacity allocation strategy for the Economy class as *penalty-averse*, as the IaaS service provider chooses to make α not too big to make sure that on average no financial penalty would be paid. Notice that α_1^* is the largest such α to also maximize usage of the reclaimed resources by the Economy class and thus maximize the revenue generated by the Economy class, which is higher than that of the Opportunistic class per unit of resources consumed. It is worth pointing out that while α_1^* appears not to explicitly depend on important parameters such as V_e and X, as later presented in (13), whether the cloud service provider should adopt a penalty-averse strategy (i.e., setting $\alpha = \alpha_1^*$) or penalty-preference strategy (i.e., setting $\alpha = \alpha_2^*$) depend on these parameters.

B. Penalty-preference Strategy

If $\Pr[\hat{u} \geq 0] \leq A_e$, which requires $\alpha \geq \alpha_1^*$, then on average there is *not* enough spare capacity after allocating $\alpha \times 100\%$ of the average amount of reclaimed resources to the Economy class to cushion resource usage fluctuations. As a result, the cloud service provider will pay a financial penalty to the Economy class users for SLO violation. Again, we want to emphasize that this does not mean that the cloud service provider would always pay financial penalty to a user during any billing period. It is in a probabilistic sense that the cloud service provider would pay financial penalty.

We now take the first order derivative of π_{reo} with respect to α, which is given as follows:

$$\frac{d\pi_{reo}}{d\alpha} = p_e \mu_{\bar{u}} (1 - X) - p_o \mu_{\bar{u}} \int_0^{\frac{1}{1+\alpha\mu_{\bar{u}}}} u f(u) du, \tag{8}$$

and the second order derivative of π_{reo} with respect to α is given as follows:

$$\frac{d^2\pi_{reo}}{d\alpha^2} = \frac{p_o\mu_{\bar{u}}^2}{(1+\alpha\mu_{\bar{u}})^3} f\left(\frac{1}{1+\alpha\mu_{\bar{u}}}\right) > 0. \quad (9)$$

If the first order derivative (8) is always negative, then the revenue decreases as α increases beyond α_1^*. Therefore, the optimal α in this case is also α_1^*: the penalty averse strategy.

If the first order derivative (8) is not always negative, there are then only two possibilities: either (8) is always positive, or it is first negative, then changes its sign, becomes positive, and remains positive. This is because the second order derivative (9) is always positive and thus the first order derivative is monotonically increasing. In this case, the optimal α will be 1 and we denote it as

$$\alpha_2^* = 1. \quad (10)$$

We refer to this as the *penalty-preference* capacity planning strategy for the Economy class.

C. The Optimal Capacity Planning Strategy

To determine whether the penalty-averse or the penalty-preference capacity planning strategy should be adopted for the Economy class, we need to compare the total revenue generated using $\alpha = \alpha_1^*$ and $\alpha = \alpha_2^*$.

Under the penalty-averse strategy, the cloud service provider allocates $\alpha = \alpha_1^*$ percentage of reclaimed resources to the Economy class and on average the cloud service provider would not pay financial penalty for the Economy class SLO violation. We thus obtain the average total revenue, denoted as $\pi_{reo,1}^*$, as follows:

$$\begin{aligned}\pi_{reo,1}^* &= V_r A_r + V_e A_e \left(\frac{1}{F^{-1}(A_e)} - 1\right) + p_o \\ &\quad \times \left(A_e - \frac{1}{F^{-1}(A_e)} \int_0^{F^{-1}(A_e)} u f(u) du\right) \quad (11)\end{aligned}$$

Under the penalty-preference strategy, the cloud service provider allocates all reclaimed resources to the Economy class, i.e., $\alpha = \alpha_2^* = 1$, and on average the cloud service provider would pay financial penalty for the Economy class SLO violation. The average total revenue in this case, denoted as $\pi_{reo,2}^*$, is as follows:

$$\begin{aligned}\pi_{reo,2}^* &= V_r A_r - V_e A_e \left(1 + \mu_{\bar{u}} - \frac{V_r}{V_e} - \mu_{\bar{u}} X\right) \\ &\quad - p_o(1+\mu_{\bar{u}}) \int_0^{\frac{1}{1+\mu_{\bar{u}}}} u f(u) du \\ &\quad + p_o F\left(\frac{1}{1+\mu_{\bar{u}}}\right). \quad (12)\end{aligned}$$

Therefore, when $\pi_{reo,1}^* \geq \pi_{reo,2}^*$, the optimal strategy is penalty-averse with $\alpha = \alpha_1^*$. Rearranging terms, this leads to the following condition:

$$\begin{aligned}&V_e A_e \left(\frac{1}{F^{-1}(A_e)} + \mu_{\bar{u}} X - 1 - \mu_{\bar{u}}\right) \\ > \; & p_o \left\{ F\left(\frac{1}{1+\mu_{\bar{u}}}\right) + \frac{E\left[u | 0 \leq u \leq F^{-1}(A_e)\right]}{F^{-1}(A_e)} - A_e \right. \\ & \left. -(1+\mu_{\bar{u}}) E\left[u | 0 \leq u \leq \frac{1}{1+\mu_{\bar{u}}}\right] \right\}. \quad (13)\end{aligned}$$

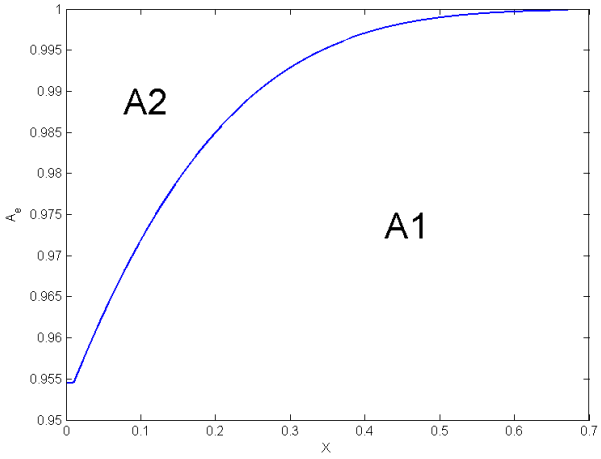

Fig. 1. The impact of the Economy class SLA parameters (A_e, X) on the Optimal Capacity Planning Strategy for the Economy Class

Notice that (13) is a complicated expression involving multiple parameters. But we see that the valuation V_r and availability SLO A_r for the Reserve class do not affect the capacity allocation strategy for the Economy class. To gain insights into the effects of these parameters on the cloud service provider's optimal Economy class capacity planning strategy, we consider a specific case by setting the distribution of resource usage rate u to a normal distribution with a mean of 0.5 and a standard deviation of 0.1, $\sim N(0.5, 0.1^2)$, and letting $V_e = 10$ and $p_o = 1$. We then vary the parameters of the SLA for the Economy class A_e and X, and plot the boundary between the penalty-averse and penalty-preference strategies. The result is shown in Figure 1.

In Figure 1, the horizontal axis is X, the financial penalty for the Economy class availability SLO violation, as a discount of the service bill. The range of X is from 0 to 0.7 and covers the common values used. For example, Amazon EC2 specifies a discount of 30% of the user's monthly bill, i.e., $X = 0.3$, when the monthly availability of On-Demand or Reserved Instances is lower than 99%. The left end of the horizontal axis thus represents very light financial penalty for the Economy class SLO violations.

The vertical axis in Figure 1 is A_e. We plot reasonable ranges of A_e for the Economy class to be useful to users, from 95% to 100%. So the upper end of the vertical axis represents higher availability SLO offered by the cloud service provider for the Economy class.

The A2 area in Figure 1 corresponds to ranges of A_e and X values that lead to a penalty-averse strategy; else, the optimal α will be 1, corresponding to a penalty-preference strategy in the A1 area. From Figure 1, we see that for given IaaS service market conditions in terms of the valuation of the Economy class V_e, the price of the Opportunistic class p_o, and the workload characteristics of the data center in terms of the distribution function of the resource usage rate u, the SLA parameters for the Economy class affect the capacity planning strategy in the following ways:

- For a given Economy class availability SLO A_e, when the penalty for SLO violation is sufficiently large, the cloud service provider will adopt a penalty-averse capacity allocation strategy to maximally offer the Economy class service without incurring the SLO violation penalty on average.
- For a given Economy class SLO violation penalty X, when the availability SLO A_e is high enough, the cloud service provider will adopt a penalty-preference capacity allocation strategy to maximally offer the Economy class service while incurring the SLO violation penalty on average.

We see that these observations are very reasonable. When the SLA sets the SLO too high, the price of a conservative Economy class capacity allocation strategy would be too high and decrease the economic benefit of the Economy class. In that case, it is better off for the cloud service provider to pay the penalty while encouraging maximum usage of the Economy class service. On the other side, if the penalty is too small, the additional revenue from the Economy class would also encourage the cloud service provider to prefer paying for SLO violation penalty to maximally realizing the revenue benefit of the Economy class.

While these qualitative observations are quite intuitive, the threshold itself is quite complex from the conditions specified by (13). This demonstrates the value of our model and analysis. We next proceed to analyze the actual revenue impact of adding the Economy class to the hierarchy of service classes offered by a cloud service provider.

V. Profitability analysis

In the previous section, we derive the optimal Economy class capacity planning strategy and pricing. However, it does not guarantee that adding this Economy class to the offering of the cloud service provider would lead to revenue increase. This is because the addition of the Economy class takes at least some reclaimed resources away from the Opportunistic class, which generates revenue without the need to pay any financial penalty for SLO violations. In this section, we compare the optimal revenue obtained for model (1) without the Economy class against the optimal revenue obtained for model (4) with the Economy class to derive conditions that guarantee the profitability of the Economy class.

First notice that the optimal total revenue of the service provider when only offering the Reserved and Opportunistic service classes can be rewritten as

$$\pi^*_{ro} = V_r A_r + p_o \left(1 - \int_0^1 u f(u) du\right). \tag{14}$$

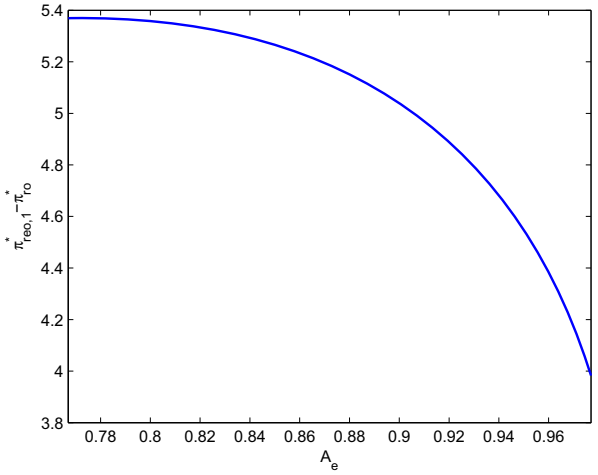

Fig. 2. The impact of the Economy class SLA parameters A_e on the Revenue Difference $\pi^*_{reo,1} - \pi^*_{ro}$

When the cloud service provider also offers the Economy class, we first consider the penalty-averse strategy. In that case, we have $\Pr[1-(1+\alpha\mu_{\bar{u}})u \geq 0] \geq A_e$ and the optimal total revenue is $\pi^*_{reo,1}$ as given in (11). We thus have the revenue difference $\pi^*_{reo,1} - \pi^*_{ro}$ after adding the Economy class using a penalty-averse strategy as

$$\begin{aligned}\pi^*_{reo,1} - \pi^*_{ro} &= V_e A_e \left(\frac{1}{F^{-1}(A_e)} - 1\right) \\ &\quad - p_o \left[1 + \frac{1}{F^{-1}(A_e)}\tilde{\mu}_u - A_e - \mu_u\right],\end{aligned}$$

where $\tilde{\mu}_u = \int_0^{F^{-1}(A_e)} u f(u) du \in (0, \mu_u)$. Therefore, if

$$p^*_e = V_e A_e > \frac{\frac{\tilde{\mu}_u}{F^{-1}(A_e)} + 1 - A_e - \mu_u}{\frac{1}{F^{-1}(A_e)} - 1} p_o, \tag{15}$$

then adding the Economy class will lead to a increase in total revenue. Let the distribution of resource usage rate u follow a normal distribution with a mean of 0.5 and a standard deviation of 0.1, $\sim N(0.5, 0.1^2)$, and let $Ve = 10$, and $p_o = 1$. In this case, A_e can vary from 0.767 to 0.977 according to $\Pr[1-(1+\alpha\mu_{\bar{u}})u \geq 0] \geq A_e$ and (15). We then plot A_e against $\pi^*_{reo,1} - \pi^*_{ro}$ in Figure 2, which shows that the revenue difference is monotonically decreasing in A_e.

In the penalty-preference case when $\Pr[1-(1+\alpha\mu_{\bar{u}})u \geq 0] < A_e$, the optimal α is equal to 1 and its corresponding optimal total revenue is $\pi^*_{reo,1}$ as given in (12). Now the revenue change after adding the Economy model $\pi^*_{reo,2} - \pi^*_{ro}$ is given by

$$\begin{aligned}\pi^*_{reo,2} - \pi^*_{ro} &= V_e A_e \left(1 + \mu_{\bar{u}} - \frac{V_r}{V_e} - \mu_{\bar{u}} X\right) \\ &\quad - p_o \left[(1+\mu_{\bar{u}})\hat{u}_u + \mu_{\bar{u}} - F\left(\frac{1}{1+\mu_{\bar{u}}}\right)\right],\end{aligned}$$

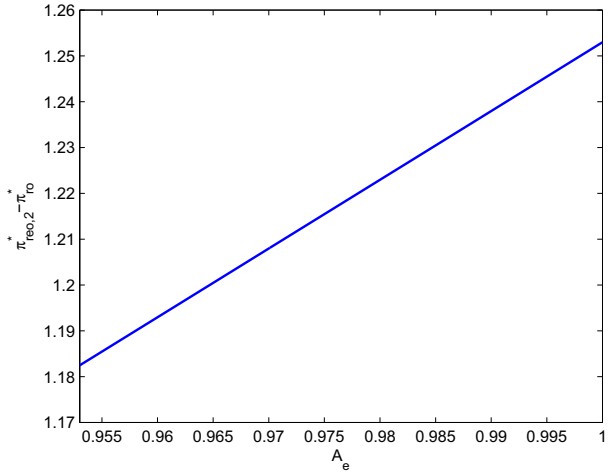

Fig. 3. The impact of the Economy class SLA parameters A_e on the Revenue Difference $\pi^*_{reo,2} - \pi^*_{ro}$

where $\hat{\mu}_u = \int_0^{\frac{1}{1+\mu_{\bar{u}}}} uf(u)du \in (0, \mu_u)$. Therefore, if

$$p_e^* = V_e A_e > \frac{(1+\mu_{\bar{u}})\hat{u}_u + \mu_{\bar{u}} - F\left(\frac{1}{1+\mu_{\bar{u}}}\right)}{1+\mu_{\bar{u}} - \frac{V_r}{V_e} - \mu_{\bar{u}} X} p_o, \qquad (16)$$

then the total revenue with the addition of the Economy class is larger than without the Economy class. Besides the parameters set for plotting Figure 3, we set $V_r = 12$ and $X = 0.3$ additionally in this case, and A_e can vary from 0.953 to 1 according to $\Pr[1-(1+\alpha\mu_{\bar{u}})u < 0] \geq A_e$ and (16). We then plot A_e against $\pi^*_{reo,2} - \pi^*_{ro}$ in Figure 3, which shows that the revenue difference is monotonically increasing in A_e.

The conditions given in (15) and (16) demonstrate the competition between the Economy class and the Opportunistic class. In order for the Economy service class to generate increase in total revenue, the valuation of this service class must be high enough to support a price p_e^* for the Economy class service that is higher than the price of the Opportunistic class multiplied by a factor that depends on various parameters affecting the Economy class.

VI. Conclusion

In this paper, we present a mathematical and economic model to analyze the introduction of a new Economy IaaS cloud service class using reclaimed cloud computing resources as proposed in [1]. Current practice of offering Opportunistic service class using reclaimed resources does not provide any SLO for the service. In contrast, the Economy class is able to offer long-term availability SLOs and makes it an attractive service class for users who need good long-term SLOs for their applications but also needs to reduce the cost they pay for cloud computing service.

By analyzing two revenue maximization models subject to user self-selection constraints, we are able to show that the optimal capacity allocation strategy for the Economy class is either penalty-averse or penalty-preference under an SLA structure used in [1]. While the threshold of the Economy class SLA parameters between these two strategies is far from trivial, we are able to clearly observe the qualitative impact of the Economy class SLA parameters on the cloud service provider's capacity planning strategy.

When the Economy class service availability SLO is set too high or the financial penalty for SLO violation is set too low, it is optimal for the cloud service provider to adopt a penalty preference strategy by offering all reclaimed resources to the Economy class users. While this strategy is optimal from a total revenue perspective, it leads to frequent SLO violations and may hurt the valuation of the Economy service class by users in the long run. In our ongoing research, we are exploring an alternative SLA design that not only optimizes the cloud service provider's total revenue, but maintains a more appropriate balance between maximizing the usage of the Economy class and maintaining the actual service quality of this new cloud computing service class.

We also derive conditions under which the Economy class will increase the total revenue for the cloud service provider. While the conditions are quite complex, the message is very clear. The valuation of the new Economy class must be sufficiently high to support an optimal Economy class pricing that can more than offset the Opportunistic class revenue dilution as a result of directing reclaimed resources from the Opportunistic class to the Economy class.

In the future, we will study extend the analysis to incorporate more realistic scenarios. In this paper, we assume that there is sufficient demand to consume resources, as was assumed in [1]. This assumption allows the use of a customer self-selection model to analyze the economic impact of the Economy cloud service class. In the future, we will relax this assumption and adopt a price-demand response approach, where the actual consumption of resources decreases as the price increases. In this paper, following [1], we assume there is only one VM configuration offered by the cloud system. We will pursue extensions to allow two or more VM configuration types in the analysis in future work. Another direction of future work is to empirically evaluate the analysis results using real-life data center traces.

Acknowledgment

Jie Xu's research was supported in part by the 2013 Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities. The authors would also like to thank Dr. Daniel Menasce, Dr. John Wilkes, Dr. Walfredo Cirne, and Mr. Marcus Carvalho for valuable feedback.

References

[1] M. Carvalho, W. Cirne, F. Brasileiro, and J. Wilkes, "Long-term slos for reclaimed cloud computing resources," in *Proceedings of the ACM Symposium on Cloud Computing*. ACM, 2014, pp. 1–13.

[2] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouillet, and D. Pendarakis, "Efficient resource provisioning in compute clouds via vm multiplexing," in *Proceedings of the 7th international conference on Autonomic computing*. ACM, 2010, pp. 11–20.

[3] B. B. Nandi, A. Banerjee, S. C. Ghosh, and N. Banerjee, "Stochastic vm multiplexing for datacenter consolidation," in *Services Computing (SCC), 2012 IEEE Ninth International Conference on*. IEEE, 2012, pp. 114–121.

[4] Amazon EC2 service level agreement, http://aws.amazon.com/ec2/sla/, accessed: 2015-05-22.

[5] P. Marshall, K. Keahey, and T. Freeman, "Improving utilization of infrastructure clouds," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*. IEEE, 2011, pp. 205–214.

[6] Amazon EC2 spot instances, http://http://aws.amazon.com/ec2/purchasing-options/spot-instances/, accessed: 2015-05-22.

[7] M. Guazzone, C. Anglano, and M. Canonico, "Energy-efficient resource management for cloud computing infrastructures," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 424–431.

[8] B. Guenter, N. Jain, and C. Williams, "Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 1332–1340.

[9] M. S. Ilyas, S. Raza, C.-C. Chen, Z. A. Uzmi, and C.-N. Chuah, "Red-bl: energy solution for loading data centers," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2866–2870.

[10] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[11] H. Viswanathan, E. K. Lee, I. Rodero, D. Pompili, M. Parashar, and M. Gamell, "Energy-aware application-centric vm allocation for hpc workloads," in *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 890–897.

[12] L. Liu, J. Xu, H. Yu, L. Li, and C. Qiao, "A novel performance preserving vm splitting and assignment scheme," in *Communications (ICC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4215–4220.

[13] C.-H. Hsu, S.-C. Chen, C.-C. Lee, H.-Y. Chang, K.-C. Lai, K.-C. Li, and C. Rong, "Energy-aware task consolidation technique for cloud computing," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 115–121.

[14] A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*. IEEE, 2010, pp. 577–578.

[15] T. Lu, M. Chen, and L. L. Andrew, "Simple and effective dynamic provisioning for power-proportional data centers," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 1161–1171, 2013.

[16] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 5, pp. 1378–1391, 2013.

[17] X. León and L. Navarro, "Limits of energy saving for the allocation of data center resources to networked applications," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 216–220.

[18] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services." in *NSDI*, vol. 8, 2008, pp. 337–350.

[19] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1. ACM, 2005, pp. 303–314.

[20] D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Resource pool management: Reactive versus proactive or lets be friends," *Computer Networks*, vol. 53, no. 17, pp. 2905–2922, 2009.

[21] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," in *ACM SIGOPS Operating Systems Review*, vol. 35, no. 5. ACM, 2001, pp. 103–116.

[22] H. Yu, V. Anand, C. Qiao, H. Di, and X. Wei, "A cost efficient design of virtual infrastructures with joint node and link mapping," *Journal of Network and Systems Management*, vol. 20, no. 1, pp. 97–115, 2012.

[23] G. Sun, H. Yu, V. Anand, and L. Li, "A cost efficient framework and algorithm for embedding dynamic virtual network requests," *Future Generation Computer Systems*, vol. 29, no. 5, pp. 1265–1277, 2013.

[24] S. Di, D. Kondo, and W. Cirne, "Host load prediction in a google compute cloud with a bayesian model," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press, 2012, p. 21.

[25] 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*. IEEE, 2010, pp. 9–16.

[26] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, 2011, p. 5.

[27] R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, "Prediction intervals for exponential smoothing using two new classes of state space models," *Journal of Forecasting*, vol. 24, no. 1, pp. 17–37, 2005.

[28] G. Gursun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 16–20.

[29] D. Niu, B. Li, and S. Zhao, "Understanding demand volatility in large vod systems," in *Proceedings of the 21st international workshop on Network and operating systems support for digital audio and video*. ACM, 2011, pp. 39–44.

[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. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: insights from google compute clusters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[32] Q. Zhang, J. L. Hellerstein, and R. Boutaba, "Characterizing task usage shapes in googles compute clusters," in *Large Scale Distributed Systems and Middleware Workshop (LADIS11)*, 2011.

[33] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *Proceedings of the Third ACM Symposium on Cloud Computing*. ACM, 2012, p. 7.

[34] Amazon EC2 Instances, http://aws.amazon.com/ec2/instance-types/, accessed: 2015-05-22.

[35] C. A. Waldspurger, "Memory resource management in vmware esx server," *ACM SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 181–194, 2002.

[36] K. S. Moorthy and I. P. Png, "Market segmentation, cannibalization, and the timing of product introductions," *Management Science*, vol. 38, no. 3, pp. 345–359, 1992.