Characterizing the Impact of the Workload on the Value of Dynamic Resizing in Data Centers

Kai Wang
Tsinghua University
Beijing, China
kaiphd@caltech.edu

Minghong Lin
California Institute of Technology
Pasadena, CA, USA
mhlin@caltech.edu

Florin Ciucu
T-Labs / TU Berlin
Berlin, Germany
florin@net.t-labs.tu-berlin.de

Adam Wierman
California Institute of Technology
Pasadena, CA, USA
adamw@caltech.edu

Chuang Lin
Tsinghua University
Beijing, China
chlin@tsinghua.edu.cn

ABSTRACT

Energy consumption imposes a significant cost for data centers; yet much of that energy is used to maintain excess service capacity during periods of predictably low load. Resultantly, there has recently been interest in developing designs that allow the service capacity to be dynamically resized to match the current workload. However, there is still much debate about the value of such approaches in real settings. In this paper, we show that the value of dynamic resizing is highly dependent on statistics of the workload process. In particular, both slow time-scale non-stationarities of the workload (e.g., the peak-to-mean ratio) and the fast time-scale stochasticity (e.g., the burstiness of arrivals) play key roles. To illustrate the impact of these factors, we combine optimization-based modeling of the slow time-scale with stochastic modeling of the fast time scale. Within this framework, we provide both analytic and numerical results characterizing when dynamic resizing does (and does not) provide benefits.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling techniques

Keywords

Data Centers, Dynamic Resizing, Energy Efficient IT

1. INTRODUCTION

Energy costs represent a significant, and growing, fraction of a data center's budget. Hence there is a push to improve the energy efficiency of data centers, both in terms of the components (servers, disks, network) and the algorithms. One specific aspect of data center design that is the focus of our work is dynamically resizing the service capacity of the data center so that during periods of low load some servers are allowed to enter a power-saving mode (e.g., go to sleep or shut down).

The potential benefits of this dynamic resizing have been a point of debate in the community [3, 1, 5]. On one hand, it is clear that, because data centers are far from perfectly energy proportional, significant energy is used to maintain excess capacity during periods of predictably low load when there is a diurnal workload with a high peak-to-mean ratio. On the other hand, there are also significant costs to dynamically adjusting the number of active servers. These costs come in terms of the engineering challenges in making this possible, as well as the latency, energy, and wear-and-tear costs of the actual "switching" operations involved.

SIGMETRICS'12, June 11–15, 2012, London, England, UK.
ACM 978-1-4503-1097-0/12/06.

The challenges for dynamic resizing highlighted above have been the subject of significant research. At this point, many of the engineering challenges associated with facilitating dynamic resizing have been resolved. Additionally, the algorithmic challenge of deciding, without knowledge of the future workload, whether to incur the significant "switching costs" associated with changing the available service capacity has been studied in depth and a number of promising algorithms have emerged.

However, despite this body of work, the question of characterizing the potential benefits of dynamic resizing has still not been properly addressed. Providing new insight into this topic is the goal of the current extended abstract. *For the full version of the work summarized here please refer to [6].*

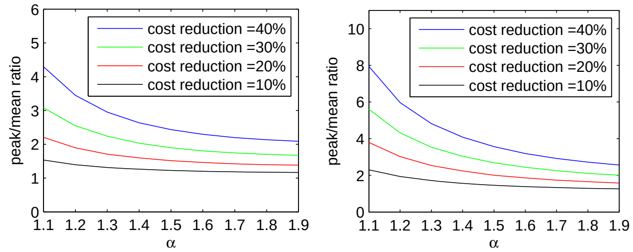
2. METHODOLOGY

The perspective of this work is that, apart from engineering challenges, the key determinant of whether dynamic resizing is valuable is the workload, and that proponents on different sides tend to have different assumptions in this regard. In particular, a key observation, which is the starting point for our work, is that there are two factors of the workload which provide dynamic resizing potential savings:

(i) Non-stationarities at a slow time-scale, e.g., diurnal workload variations.
(ii) Stochastic variability at a fast time-scale, e.g., the burstiness of request arrivals.

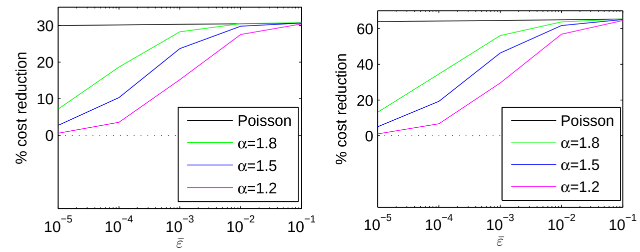
The goal of this work is to investigate the impact of and interaction between these two features with respect to dynamic resizing.

To this point, we are not aware of any work characterizing the benefits of dynamic resizing that captures both of these features. There is one body of literature which provides algorithms that take advantage of (i), e.g., [4]. This work tends to use an optimization-based approach to develop dynamic resizing algorithms. There is another body of literature which provides algorithms that take advantage of (ii), e.g., [2]. This work tends to assume a stationary



(a) Hotmail (b) MSR

Figure 1: Characterization of the cost reduction of optimal dynamic resizing under different levels of burstiness, α, and peak-to-mean ratio in the workload.



(a) Hotmail (b) MSR

Figure 2: Characterization of the impact of the SLA, $\bar{\varepsilon}$, and burstiness, α, on the cost reduction achieved by optimal dynamic resizing.

queueing model with Poisson arrivals to develop dynamic resizing algorithms.

The first contribution of our work is to provide an analytic framework that captures both effects (i) and (ii). We accomplish this by using an optimization framework at the slow time-scale, which is similar to that of [4], and combining this with stochastic network calculus and large deviations modeling for the fast time-scale, which allows us to study a wide variety of underlying arrival processes. We consider both light-tailed models and heavy-tailed models with various degrees of burstiness by varying the tail index α which describes the shape of the tail [1].

The interface between the fast and slow time-scale models happens through a constraint in the optimization problem that captures the Service Level Agreement (SLA) for the data center, which is used by the slow time-scale model but calculated using the fast time-scale model:

$$\mathbb{P}(D_k > \bar{D}) \leq \bar{\varepsilon} \ , \qquad (1)$$

where we use D_k to represent the steady-state delay during frame k, and $(\bar{D}, \bar{\epsilon})$ to represent an SLA of the form "the probability of a delay larger than $\bar{D}$ must be bounded by probability $\bar{\varepsilon}$".

3. RESULTS

Using this modeling framework, we are able to provide both analytic and numerical results that yield new insight into the potential benefits of dynamic resizing. Specifically, we use trace-driven numerical simulations to study (i) the role of burstiness for dynamic resizing, (ii) the role of the peak-to-mean ratio for dynamic resizing, (iii) the role of the SLA for dynamic resizing, and (iv) the interaction between (i), (ii), and (iii). The key realization is that each of these parameters are extremely important for determining the value of dynamic resizing. In particular, for any fixed choices of two of these parameters, the third can be chosen so that dynamic resizing does or does not provide significant cost savings for the data center. Thus, performing a detailed study of the interaction of these factors is important. To that end, we provide concrete illustrations of which settings of peak-to-mean ratio, burstiness, and SLAs dynamic resizing are and are not valuable (e.g., Figure 1). Hence, debate about the *potential* value of dynamic resizing can be transformed into debate about characteristics of the workload and the SLA.

There are some interesting facts about these parameters individually that our case studies uncover. Two important examples are the following (see Figure 2). First, while one might expect that increased burstiness provides increased opportunities for dynamic resizing, it turns out the burstiness at the fast time-scale actually reduces the potential cost savings achievable via dynamic resizing. The reason is that dynamic resizing necessarily happens at the slow time-scale, and so the increased burstiness at the fast time-scale actually results in the SLA constraint requiring *more* servers to be used at the slow time-scale due to the possibility of a large burst occurring. Second, it turns out the impact of the SLA can be quite different depending on whether the arrival process is heavy- or light-tailed. In particular, as the SLA becomes more strict, the cost savings possible via dynamic resizing under heavy-tailed arrivals decreases quickly; however, the cost savings possible via dynamic resizing under light-tailed workloads is unchanged.

In addition to detailed case studies, we provide analytic results that support many of the insights provided by the numerics. In particular, we prove theorems to provide monotonicity and scaling results for dynamic resizing in the case of Poisson arrivals and heavy-tailed, self-similar arrivals.

4. ACKNOWLEDGMENTS

This research is supported by the 973 Program of China (No. 2010CB328105), the National Natural Scientific Foundation of China (No. 61020106002 and No.60973107), and NSF grant CNS 0846025 and DoE grant DE-EE0002890.

5. REFERENCES

[1] 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 *Proceedings of the 5th USENIX NSDI*, pages 337–350, 2008.

[2] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch. Optimality analysis of energy-performance trade-off for server farm management. *Performance Evaluation*, 67(11):1155–1171, 2010.

[3] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel. The cost of a cloud: research problems in data center networks. *SIGCOMM Comput. Commun. Rev.*, 39:68–73, December 2008.

[4] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *Proceedings IEEE INFOCOM*, pages 1098–1106, 2011.

[5] D. Meisner, C. M. Sadler, L. A. Barroso, W.-D. Weber, and T. F. Wenisch. Power management of online data-intensive services. In *Proceedings of the 38th ACM International Symposium on Computer Architecture*, 2011.

[6] K. Wang, M. Lin, F. Ciucu, A. Wierman, and C. Lin. Characterizing the impact of the workload on the value of dynamic resizing in data centers. Technical report, 2012.

[1] A concrete example is generating jobs in every slot according to i.i.d. Pareto random variables X_i with tail distribution for all $x \geq b$: $P(X_i > x) = (x/b)^{-\alpha}$, in which smaller values of α indicate heavier tails and, thus, more burstiness.