Estimating Optimal Cost of Allocating Virtualized Resources with Dynamic Demand

Haiyang Qian and Deep Medhi
University of Missouri-Kansas City

***Abstract*—Considering the dynamics of the demand on virtulized resources is indispensable to reduce the operational costs in the context of "pay-as-you-go" model. In this paper, we formulate the optimization problem of minimizing the operational costs with dynamic demand. Furthermore, a new simple index to estimate the optimum is proposed.**

I. Introduction

Data centers are designed to provision sufficient virtualized resources at all times to meet the demands of users. Nowadays the users who have "elastic demand" take advantage of the "pay-as-you-go" model and only request needed demands which vary over time. Therefore, data centers must have the capability to satisfy the "peak" requirements. On the other hand, keeping all servers running at maximum capacity all times wastes energy since most of the time is "off-peak". Consolidating virtual machines dispersed across the servers in the data center into a smaller number of servers by migrating and turning unnecessary servers off, can reduce energy consumption. However, turning server on and off causes wear and tear costs and migrating virtual machines impose management overhead to the system. Many processors have so called Dynamic Voltage/Freqeuency Scaling (DVFS) capacity that allows processors to scale the frequency to reduce the power consumption. In addition, the dynamics of demand (temporal variation, uncertainty, etc.) impose unique challenges to this problem. We formulate this problem by mathematical programming. Related work on formulating the optimization problem can be found in [7], [3], [4], [1], [2], [8], [6]. We propose a simple index to estimate the optimum by synthesizing the characteristics of the demand and our optimization framework. To the best of our knowledge, this is the first work on this topic.

II. Formulation

The data center has I servers. Let $\mathcal{I}$ denote the set of the servers. Let $\mathcal{J}(i)$ be the frequency option set for server i. In a homogeneous server cluster case, $\mathcal{J}(i) = \mathcal{J}, \forall i$. There are J frequency options in $\mathcal{J}$. Server i running at the j-th frequency option can offer a capacity of V_{ij} while satisfying the SLA. The power consumption of running server i at j-th frequency is denoted by C_{ij} per time unit. The wear and tear costs of turning a server on and off is denoted by C_s^+, C_s^-, respectively. We divide the planning period Υ hours into T equal time slots and the *duration of a time slot (slot size)* is then $\tau = \Upsilon/T$ hours (usually, we refer to the slot size in minutes). Let $\mathcal{T}$ be the set of these slots. At the review point (starting point of each time slot), the number of active servers and their frequencies is configured. Let binary decision variables $y_{ij}(t)$ denote if server i is running at frequency option j at time slot t.

There are a couple of constraints in this problem. A server can only be operated at one frequency in a time slot:

$$\sum_{j\in\mathcal{J}} y_{ij}(t) \leq 1,\ \forall i \in \mathcal{I}, \forall t \in \mathcal{T} \tag{1}$$

The second constraint is the demand requirement. Let the demand on CPU at time t be $D(t)$. First we assume the demand can be predicted deterministically:

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} V_{ij}\cdot y_{ij}(t) \geq D(t),\ \forall t \in \mathcal{T} \tag{2}$$

The demand can be predicted *probabilistically* rather than deterministically. Namely, the demand is a random variable, denoted by $\widetilde{D}(t)$. Based on SLA, we guarantee the probability of the provided capacity being no less than the demand is no less than a certain percentage based. The probability of demand being no less than $\widetilde{D}(t)$ is no less than p at time slot t is

$$P\big(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} V_{ij}\cdot y_{ij}(t) \geq \widetilde{D}(t)\big) \geq p,\ \forall t \in \mathcal{T} \tag{3}$$

If the cumulative distribution function (CDF) of demand at time slot t is given by F_t,(3) is equivalent to

$$\sum_{i\in I}\sum_{j\in J} V_{ij}\cdot y_{ij}(t) \geq F_t'(p),\ \forall t \in \mathcal{T} \tag{4}$$

where F_t' denotes the inverse function of F_t. Note that this model is very general. An example that this model can be applied to is described as follows. The demand is forecasted by an expected value with certain fluctuation. The actual demand is uniformly distributed among this range.

Our objective is to minimize the operational costs of the running servers:

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} C_{ij}\cdot y_{ij}(t) + \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}\Big(C^+\cdot y_i^+(t) + C^-\cdot y_i^-(t)\Big) \tag{5}$$

where binary variables, $y_i^+(t)$ and $y_i^-(t)$, represent turning on/off at the review point of time slot t. $y_i^+(t) = 1$ means the server i is turned on at time t. Conversely, $y_i^-(t) = 1$ means the server i is turned off at time t. 0 indicates no change

978-0-9836283-0-9

This paper was peer reviewed by subject matter experts for publication in the Proceedings of ITC 2011

from time slot $t-1$ to t. The purpose of introducing these two variables is to transform the original quadratic objective function to a linear objective function. For the details of this transformation and other constraints associated with it, refer to our early paper [9].

III. Estimating Optimum

Within this optimization framework, is it possible to use a certain index to estimate the optimal cost before actually executing the optimization given the demand? We introduce the *demand dent* to solve this problem. In our preliminary work, we do not consider DVFS. We have the following key observations on how to get optimum based on our optimization framework: 1) We must turn on more servers if the capacity is not enough. 2) We may or may not turn off servers if the capacity is superfluous. We choose whichever can achieve global optimum. Namely, we choose the minimum between wasting energy and introducing reallocation overhead. To accomplish this idea, we define the *demand dent* as the demand slot(s) (in the demand time series) inclusively between t and t' (where $t' > t$) where $D(t) < D(t-1)$ && $D(t'+1) > D(t')$. Let $\mathcal{K}$ denote the set of the demand dents. Define the summation of the minimum between wasted energy and reallocation overhead in each demand dent as *demand dent index*, (I_D):

$$I_D = \sum_{k \in \mathcal{K}} \Big(\min\big(\sum_{k=t}^{t'} C(D(t-1) - D(t)), \\ C^+(D(t-1) - D(t)) + C^-(D(t'+1) - D(t')) \big) \Big)$$

where C is the unit power price.

IV. Preliminary Results

In our study, we consider a server cluster of 100 identical servers. The CPU frequency set and power consumptions are adopted from [4] except that we use the maximum frequency only. For ease of computation, the capacity of each server is normalized to 1. In [5], Greenberg *et. al.* used $.07 per KWH as the utility price. We use the same utility price. The cost of turning on is 0.32 cents and the cost of turning off is 0.205 cents. [9] describes the details of deriving cost parameters. We assume that the utilization of the data center is 20%. Thus the average demand is 20. We also assume that the demand profile is forecasted and profiled every 5 minutes. Due to the diurnal behavior associated with human beings' working cycles, we chose the 8 hour work time as the planning horizon where the dynamically changing demand from one time slot to another is generated for our study. In our evaluation, the demand at each time slot is deterministically given and $\alpha = 0$. We generate 6 different distributions with the same average of 20. Each distribution is generated 5 times with 5 different seeds. These distributions include smooth (Erlang), exponential, and bursty (Hyper-exponential) distributions. The sorted versions of three distributions give the best optimum for the same distribution. (We use this to illustrate that squared coefficient of variance cannot do the job.) We ran the optimization model using CPLEX through its integer programming solver on an Intel(R) Pentium(R)IV 3.00GHz with 2GB memory. Through preliminary runs, we observed that the overall cost does not improve if we allow 2,000 branch and cut nodes in CPLEX; thus, in our study, we set the branch and cut nodes limit to 2,000.

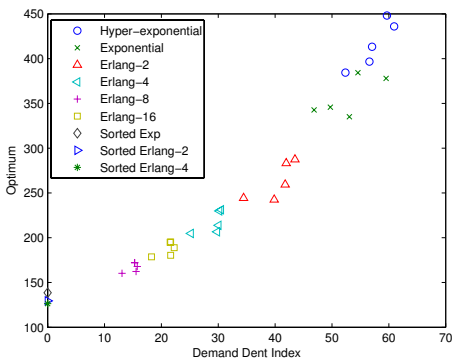

Fig. 1. **Demand Dent Index Versus Optimum**

As we can see in Fig. 1, the demand dent index is almost linearly correlated to the optimum. Thus the demand dent is a good indicator of the optimal cost yielded by our optimization framework.

V. Conclusion and Future Works

We construct mathematical programming models to optimize the virtualized resource allocation with dynamic demand. The probabilistic aspect of the demand is considered here. We are able to estimate the optimum by using the demand dent index. We are using regression methods to parameterize the relationship between the optimum and the demand dent index. As to the future, we will complete the demand dent index by including DVFS and demand with uncertainty.

References

[1] L. Bertini, J. C. B. Leite, and D. Mossé, "Power optimization for dynamic configuration in heterogeneous web server clusters," *J. Syst. Softw.*, vol. 83, no. 4, pp. 585–598, 2010.
[2] R. Bianchini and R. Rajamony, "Power and energy management for server systems," *IEEE Computer*, vol. 37, p. 2004, 2004.
[3] M. Bichler, T. Setzer, and B. Speitkamp, "Capacity planning for virtualized servers," in *Workshop on Information Technologies and Systems (WITS)*, Milwaukee, Wisconsin, USA, 2006.
[4] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," *SIGMETRICS Perform. Eval. Rev.*, vol. 33, no. 1, pp. 303–314, 2005.
[5] 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.*, vol. 39, no. 1, pp. 68–73, 2009.
[6] L. A. Johnson and D. C. Montgomery, *Operations Research in Production Planning, Scheduling, and Inventory Control*. Wiley, 1974.
[7] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Dynamic cluster reconfiguration for power and performance," 2002.
[8] M. Pióro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann Publishers, 2004.
[9] H. Qian and D. Medhi, "Server operational cost optimization for cloud computing service providers over a time horizon," in *USENIX Hot'ICE*, Boston, MA, USA, 2011.