Journal of Network and Computer Applications ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Energy efficiency aware load distribution and electricity cost volatility control for cloud service providers

Debdeep Paul [a,*], Wen-De Zhong [a], Sanjay K. Bose [b]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore
[b] Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati 781039, India

ARTICLE INFO

Keywords:
Energy efficiency
Electricity markets
Internet data centers
Online algorithms
Model predictive control

ABSTRACT

This paper consider the case of a cloud service provider (CSP) who owns multiple geographically distributed data centers, with collocated sources of renewable energy. We investigate load distribution strategies to minimize electricity cost and increase renewable incorporation subject to compliance with service level agreement (SLA), considering the adverse effects of switching the servers. Our work provides some insights on the performance of different algorithms for geographical load balancing (GLB) in terms of electricity cost, renewable energy integration and number of server switching. Our proposed strategies incorporate a new way of capturing the server switching cost. We show that, instead of modeling switching cost through a linear function, the proposed technique of modeling switching cost through variance achieves a better tradeoff between some important parameters. Since the three major input parameters-electricity price, renewable energy and number of job requests-vary over time, the average cost of electricity per job request may also exhibit dramatic fluctuations. We propose to tackle this volatility by controlling the average cost of electricity per job request through leveraging contracts in the forward electricity market, and determine the optimal amount of electricity to be procured in the forward electricity market. We show that our proposed strategy substantially reduces the variance of the average cost of electricity per job and that this price risk mitigation is achieved with a decrease in the cumulative electricity cost.

1. Introduction

The past few decades have witnessed a tremendous growth in the various services provided online through the Internet. This is continuing with the advent of innovative technologies such as Mobile Computing (Satyanarayanan et al., 2009), Internet of Things (IoT) (Tang et al., 2014), Big Data (Lynch, 2008) and other emerging applications. Contrary to traditional systems, these days a new competitor does not have to incur a huge capital expenditure (Capex) to enter into the market, since most of the services are available from the existing cloud service providers (CSPs) in a pay-as-you-go manner. To support the ever increasing customer demands for IT services, today's CSPs incur a huge electricity cost to support their day to day operations. In fact it is reported that a significant proportion of the total operational expenditure (Opex) of a typical CSP goes towards electricity costs (Greenberg et al., 2008). As per an official report, Google alone consumed $138M worth of electricity (Qureshi et al., 2009; Glanz, 2011) in 2011; the cost today must be substantially higher. Typically, in data centers resources are provisioned by conservative estimates and result in substantial over provisioning. At different granularity levels within a data center such as at cluster level, at rack level and at an aggregated level the resources provisioned are highly under-utilized. It turns out that the over-provisioning increases when we go up in the hierarchy and consider a bigger system (Fan et al., 2007). Indeed, studies have found that the overall utilization of the total capacity in data centers can be as low as 20% (Barroso and Holzle, 2007; Armbrust et al., 2010). An important reason why this happens is because the service demands (job requests) vary dramatically over time. It may be noted that under-provisioning of resources will result in violation of service level agreement (SLA) (most commonly reflected through response time), and will eventually have a detrimental effect on business revenue of the CSP. As an instance, Amazon reported that it tends to incur a loss of 1% of sales revenue for an increase of 100 ms in response time, whereas Google reported that a 500 ms increase in response time translates into a 20% reduction in the revenue (Linden, 2006). These indicate that the business revenue of a CSP is highly sensitive to the service response time, which is closely associated with

* Corresponding author. Tel.: +65 6790 4540; fax: +65 6793 3318.
E-mail addresses: debdeep1@e.ntu.edu.sg (D. Paul), ewdzhong@ntu.edu.sg (W.-D. Zhong), skbose@iitg.ernet.in (S.K. Bose).

http://dx.doi.org/10.1016/j.jnca.2015.08.012
1084-8045/

the amount of resources provisioned at a particular instant. Moreover, today's commercial servers are far from being energy proportional. They consume a fairly high proportion of peak power even when they are idle. For instance, engineers from Google reported that a commercial server typically consumes 60% of its peak power even when it is idle (Barroso and Holzle, 2007). These facts indicate that there is a lot of room for improvement in the energy consumption of massive data centers. This can be easily done by turning off the unused servers during the periods of low load while consolidating the load within the active servers. Moreover, there are several non-obvious benefits of lower server power consumption, since the reduction in computational power will proportionately reduce the power consumed by cooling system, power distribution system and other auxiliary systems as well. In this paper, we address both the financial aspect of decreasing the electricity cost incurred by the CSP and the social issue of increasing renewable penetration into the grid. This is significant because the world's data centers are estimated to consume more than 2% of the total electricity consumption (Cook, 2014). An even more alarming fact is that the demand for electricity for the cloud infrastructures is going to rise by at least 60% by 2020 (Cook, 2014). In this work, we consider the case of a CSP who owns multiple data centers at various geographic locations. The issue of load distribution between multiple geographically distributed data centers has been referred to as *geographical load balancing* (GLB) in the literature (Liu et al., 2011a, 2011b).

Specifically, this paper makes the following contributions,

- We present a holistic framework for modeling and distributing the jobs among geographically distributed data centers considering the electricity price in the wholesale market, local renewable availability, cooling related conditions (at a macroscopic level), carbon emission effects and, most crucially, the adverse effects of server switching. We use these to investigate the various trade-offs so as to have more clarity and insight on the optimal choice for GLB.
- We present, analyse and compare the relative performance between two algorithms for GLB based on Model Predictive Control (MPC), in terms of the cost of electricity, renewable integration and switching. While one of them is quite close to the existing works (Adnan et al., 2012; Lin et al., 2013), the other approach is a novel one which has not been explored as yet.
- We propose and evaluate a strategy to determine the optimal trading in the forward electricity market to deal with the issue of risk associated with the time variability in electricity price, renewables and amount of job requests.
- We present and discuss the results of extensive simulations with real data traces on electricity price, renewable generation, cooling related conditions, workload, and carbon emission effects at 10 locations across the globe, to emulate the real complexity of the system.

The rest of the paper is organized as follows. In Section 2, we discuss the related works relevant to this work. In Section 3, we present the model and propose two optimization problems with different structures of the objective function and solve them. In Section 4, we introduce a strategy to perform hedging on cost volatility of electricity by buying optimal amount of electricity in the forward market. In Section 5, we describe the simulation setup and the data traces used in our study. In Section 6, we discuss the results obtained through simulations and provide some useful insights to guide the CSPs on the deployment of an appropriate load distribution strategy. In Section 7, we present our conclusions and briefly indicate some interesting directions for future work to address meaningful issues.

2. Related work

Financial and social pressures have brought about significant improvements in decreasing the energy consumption of data centers and increasing their energy efficiency. Fairly exhaustive surveys are available in Beloglazov et al. (2012) and Pedram (2012) on the progress made to address energy consumption and efficiency issues for data centers. We leverage the spatial variation of electricity price and the availability of renewable energy to reduce the overall electricity cost and increase renewable integration into the grid. A pioneer work which came up with the idea of exploiting spatial variation of electricity price to reduce the cost of electricity incurred by the CSP is Qureshi et al. (2009). They studied temporal and geographical variation in electricity price and proposed heuristic based algorithms for distance constrained electricity price aware routing. Another work in Rao et al. (2012) with the same goal, modeled the system as a constrained mixed integer nonlinear programming problem (MINLP) and solved it using the Generalized Benders Decomposition (GBD) technique to obtain optimal load balancing and power control. Subsequently, several other works have explored the diversity in green energy availability by considering renewable energy aware load distribution along with overall cost reduction. In Liu et al. (2011a, 2011b), the authors merged the two objectives namely, electricity cost reduction and renewable energy incorporation. The work in Liu et al. (2011b) presented two distributed algorithms to achieve optimal load distribution by minimizing a linear combination of the energy cost and the revenue lost due to delay. A rigorous investigation on the feasibility of powering data centers entirely by renewable energy is presented in Liu et al. (2011a). Therein, the authors proposed that GLB can be optimally integrated with a reasonable amount of storage to achieve almost zero use of conventional brown energy from the grid. They also provided insights on the optimum portfolio of solar and wind energy to achieve this ambitious target. The work in Gao et al. (2012) tackled the issue of carbon emission and proposed algorithms for carbon footprint reduction; these have also been addressed in this paper. A good summary of the recent developments related to the integration of renewable energy in data centers is presented in Deng et al. (2014). In general, the literature on cloud resource provisioning is quite rich. The recent developments in different domains of cloud services and resource management practices can be found in Manvi and Shyam (2014) and Sun et al. (2014). Some promising and meaningful future research directions have also been presented in these works. The work in Zuo et al. (2013) presented a sophisticated method for cloud computing resource evaluation based on entropy optimization, considering the uncertainty. Another important aspect considered by us in this paper is the switching of servers and the adverse effects (cost) associated with this operation (i.e. transition from ON state to OFF state and vise versa). These tie in the problem we have considered, with the analysis of *online algorithms* and their implementation for *energy efficiency* in computer systems. The works aiming to strike a trade-off between saving energy by turning off some servers (or transitioning into low power state) during the periods of low load and switching cost (not to turn ON and OFF too frequently) with stochastic arrivals of jobs are Gandhi et al. (2010a), Lin et al. (2013) and Lu et al. (2013). A good overview and summary of notable contributions in the areas of both deterministic and stochastic algorithms focusing on energy efficiency aspects is provided in Albers (2010). This presented the performance of various types of online algorithms (with knowledge up to the present) with respect to the hypothetical *offline* version of the algorithm (with perfect future knowledge) in terms of competitive ratio.

In various different fields, scheduling plays a crucial role to achieve energy efficiency. As an instance in Xiao et al. (2010) the

Please cite this article as: Paul D, et al. Energy efficiency aware load distribution and electricity cost volatility control for cloud service providers. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.08.012