# Capping the Electricity Cost of Cloud-Scale Data Centers with Impacts on Power Markets

Yanwei Zhang, Yefu Wang, and Xiaorui Wang
University of Tennessee, Knoxville, TN 37996
{yzhang82, ywang38, xwang}@eecs.utk.edu

## ABSTRACT

In this paper, we propose a novel *electricity cost capping* algorithm that not only minimizes the electricity cost of operating cloud-scale data centers, but also enforces a cost budget on the monthly electricity bill. Our solution first explicitly models the impacts of power demands on electricity prices and the power consumption of cooling and networking in the minimization of electricity cost. In the second step, if the electricity cost exceeds a desired monthly budget due to unexpectedly high workloads, our solution guarantees the quality of service for premium customers and trades off the request throughput of ordinary customers. We formulate electricity cost capping as two related constrained optimization problems and propose an efficient algorithm based on mixed integer programming. Simulation results show that our solution outperforms the state-of-the-art solutions by having lower electricity costs and achieves desired cost capping with maximized request throughput.

## Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems

## General Terms

Management, Performance, Economics

## Keywords

Cloud-scale data centers, electricity cost, cost capping

## 1. INTRODUCTION

Minimizing the energy consumption of data centers has recently attracted a lot of research efforts. However, much less attention has been given to a related but different research topic: minimizing the electricity bill of a network of data centers by leveraging different electricity prices in different geographical locations to distribute workloads among those locations. This research topic is important for many Internet service providers, such as Google, Microsoft, and Yahoo!, to minimize their operating costs, because they commonly have massive and geographically distributed data centers to support various services such as cloud computing.

A few initial studies have been recently conducted to address the problem of electricity cost minimization [1, 2]. The key idea of those studies is to periodically monitor the time-varying electricity prices of the regions where data center sites are located. Based on the price information, Internet requests are routed to those sites where electricity prices are relatively low for minimized operating costs. While those studies have shown promise, they have an unrealistic assumption that the huge power demands of data centers have no impact on electricity prices. In other words, data centers are treated simply as *price takers* in the power markets and their electricity prices are assumed to be irrelevant to their power demands at a given time point. However, the reality in power market operation is that electricity prices are frequently adjusted mainly based on a well-known policy called the Locational Marginal Pricing (LMP) methodology [3]. According to LMP, electricity prices depend not only on geographical region and time, but also on the locational supply and demand of power. Therefore, while traditional small-scale enterprise data centers may be assumed to be passive price takers, this assumption is no longer valid for cloud-scale data centers whose sizes are much larger. For example, some data centers host more than 100,000 servers [4] and can draw tens to hundreds of megawatts of power at peak. As a result, cloud-scale data centers become the major power consumers of power suppliers and thus are now *price makers*. Therefore, the power demands of cloud-scale data centers have signficant impacts on electricity prices and the impacts must be addressed for minimized electricity costs.

*Capping the electricity cost* of could-scale data centers is another equally important issue for cloud-service providers, in addition to cost minimization. Since the electricity cost of operating data centers has become a significant portion (20% or more) of the monthly costs of those providers [4], it is a common business procedure for them to allocate a monthly budget for electricity cost. However, due to the high variations in data center workloads, it is usually difficult to enforce such a desired budget on electricity cost. For example, breaking news on major newspaper websites may incur a huge number of accesses in a short time and thus lead to unexpectedly high electricity costs for data centers. Note that cost minimization alone cannot enforce a desired electricity cost cap, because a monthly budget for electricity is commonly made based on history data with a certain safety margin. Therefore, if similar events occur frequently in a month and no effective methods are taken to control the cost, the monthly budget is likely to be violated.

To enforce a desired cost cap in the face of unexpectedly high workloads, a service provider may need to differentiate premium customers who pay for their services from ordinary customers who enjoy complimentary services. The optimization objective is to guarantee the quality of service (*e.g.*, response time) for premium customers, while reducing (to the minimum degree) the request throughput of ordinary customers for lowered electricity use and costs. Hence, we argue that electricity cost capping is becoming an increasingly important issue, as cloud-scale data centers are rapidly expanding their sizes. Cost capping should be addressed together with power capping, which is recently proposed to cap the power consumption of a single data center [5]. In order to cap the electricity use and cost of cloud-scale data centers, the power cap of each data center must first be enforced to avoid financial penalty [5]. The total electricity cost of all the distributed data centers


*HPDC'11*, June 8–11, 2011, San Jose, California, USA.
ACM 978-1-4503-0552-5/11/06.

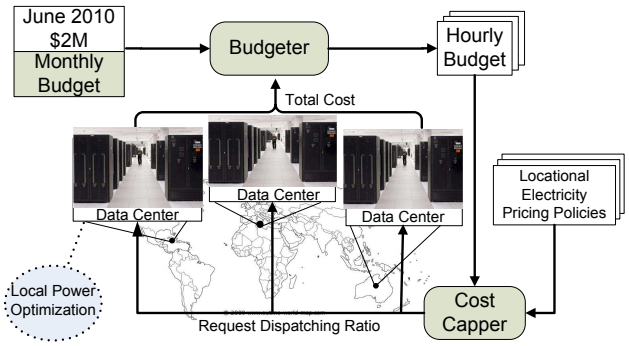


**Figure 1: Proposed electricity cost capping architecture for distributed cloud-scale data centers.**

should then be controlled to avoid resulting in a high budget deficit for the cloud-service provider. Cost capping offers cloud-service providers a flexible and effective way to achieve maximized return within their sometimes stringent budget.

In this paper, we propose a novel electricity cost capping algorithm that not only minimizes the electricity cost, but also enforces a cost budget on the monthly bill for cloud-scale data centers. In the first step, our solution explicitly models the impacts of power demands on electricity prices and the power consumption of cooling and network in the minimization of electricity cost. In the second step, if the minimized electricity cost still exceeds the desired monthly budget due to unexpectedly high workloads, our solution guarantees the quality of service for premium customers and trades off the request throughput of ordinary customers. To our best knowledge, our work presents the first study on electricity cost capping for cloud-scale data centers.

## 2. SYSTEM ARCHITECTURE

In our work we assume that a network of data centers share a cost budget in every budgeting period determined by the administrator of Internet applications. We also assume that the locational pricing policies, *i.e.*, how the changes in power consumption of data centers affect the electricity prices in local power markets, are available from ISO. As shown in Figure 1, the key components in our cost management framework include a centralized *cost capper* and *budgeter* that are invoked periodically in every *invocation period*. In this paper, we use one month as the budgeting period and one hour as the invocation period. Those invocation periods are suggested to be good trade-offs between management granularity and actuation overheads [6] for data center-level management algorithms.

When the budgeter receives a monthly budget at the beginning of the budgeting period from the system administrator, it breaks the monthly budget into hourly budgets based on the historical incoming workload data. Our budgeting algorithm is a static budgeting method, since we determine the hourly cost weight of each day based on the historical workload of the same hour in the last month. At the beginning of every invocation period, the hourly budgets of the remaining hours in the month are recomputed based on the monthly cost budget and the cost already consumed in previous invocation periods. Then, the cost capper determines the workload allocations of every data center such that:

- The total electricity cost of all the data centers is minimized and is below the budget of the current hour.
- The QoS of premium customers is guaranteed while the QoS of ordinary customers is provided in a best effort.

As discussed above, our cost capping algorithm includes two steps. (1) In the first step, the algorithm solves a ***cost minimization*** problem that minimizes the total cost of data centers with the consideration of the locational pricing policies, by distributing the Internet requests to different data centers in an efficient way. (2) In the second step, the algorithm compares the computed cost found in the first step with the given hourly budget. If the computed cost is below the budget, the workload allocations determined in the first step is enforced. Otherwise, the algorithm solves a ***throughput maximization within cost budget*** problem that determines an admission rate to enforce admission control only for requests from ordinary customers. The capping algorithm also determines the web request allocation to every data center such that the total cost of data centers is controlled below the cost budget. In addition, we assume that each data center has a local optimizer to dynamically minimize the number of active servers in the data center based on an M/M/n performance model [2], given the distributed workload.

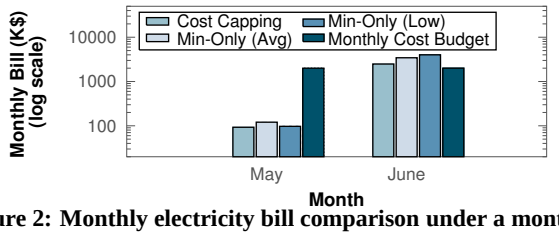


**Figure 2: Monthly electricity bill comparison under a monthly cost budget of $2*M*.**

## 3. EVALUATION RESULTS

We use a real-world web request trace file of the 1998 World Cup and a power consumption trace file from the real-world power market [7] to evaluate the proposed Cost Capping algorithm. We compare with Min-Only, a state-of-the-art cost minimization algorithm designed for Internet-scale data centers [2].

Figure 2 shows that for a relatively light workload (*e.g.*, in May), the monthly electricity costs determined by both Cost Capping and Min-Only stay within the cost budget due to the abundant cost budget. However, Cost Capping can gain 23.0% and 4.8% more electricity cost savings over two variants of the baseline: Min-Only (Avg) and Min-Only (Low), respectively. This figure also shows that in June when the requests from customers are heavy, the electricity bills determined by Min-Only (Avg) and Min-Only (Low) exceed the monthly cost budget by 71.9% and 99.5%, respectively, due to their unawareness of the stringent cost budget. Note that the monthly electricity bill by Cost Capping exceeds the given cost budget by 24.2%. This is due to the fact that we have to guarantee the QoS for premium customers despite a stringent cost budget.

More results are available in our tech report [8].

## 4. REFERENCES

[1] A. Qureshi *et al.*, "Cutting the electric bill for internet-scale systems," in *SIGCOMM*, 2009.

[2] L. Rao *et al.*, "Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment," in *INFOCOM*, 2010.

[3] F. Li, "Continuous locational marginal pricing (CLMP)," *IEEE Transactions on Power Systems*, vol. 22, no. 4, 2007.

[4] A. Greenberg *et al.*, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, 2008.

[5] X. Wang *et al.*, "SHIP: Scalable Hierarchical Power Control for Large-Scale Data Centers," in *PACT*, 2009.

[6] F. Ahmad *et al.*, "Joint optimization of idle and cooling power in data centers while maintaining response time," in *ASPLOS*, 2010.

[7] "PJM Training Materials LMP 101, PJM," http://pjm.com.

[8] Y. Zhang, Y. Wang, and X. Wang, "Electricity bill capping for cloud-scale data centers that impact the power markets, Tech Report, EECS, University of Tennessee," http://pacs.ece.utk.edu/TRcapping.pdf, 2011.