Managing the Cost, Energy Consumption, and Carbon Footprint of Internet Services*

Kien Le†, Ozlem Bilgir‡, Ricardo Bianchini†, Margaret Martonosi‡, Thu D. Nguyen†

†Department of Computer Science
Rutgers University, Piscataway, NJ
{lekien,ricardob,tdnguyen}@cs.rutgers.edu

‡Department of Electrical Engineering
Princeton University, Princeton, NJ
{obilgir,mrm}@princeton.edu

ABSTRACT

The large amount of energy consumed by Internet services represents significant and fast-growing financial and environmental costs. This paper introduces a general, optimization-based framework and several request distribution policies that enable multi-data-center services to manage their brown energy consumption and leverage green energy, while respecting their service-level agreements (SLAs) and minimizing energy cost. Our policies can be used to abide by caps on brown energy consumption that might arise from various scenarios such as government imposed Kyoto-style carbon limits. Extensive simulations and real experiments show that our policies allow a service to trade off consumption and cost. For example, using our policies, a service can reduce brown energy consumption by 24% for only a 10% increase in cost, while still abiding by SLAs.

Categories and Subject Descriptors: C.4 [Performance of System]: Modeling techniques

General Terms: Algorithms, Design, Experimentation, Performance.

1. INTRODUCTION

Data centers are major energy consumers [4]. In 2006, the data centers in the US consumed 61.4 Billion KWhs. Under current trends, this usage will nearly double by 2011 for an overall electricity cost of $7.4 Billion per year. These enormous electricity consumptions translate into large carbon footprints, since most of the electricity in the US is produced by burning coal, a carbon-intensive energy production approach. (We refer to energy produced by carbon-intensive means as "brown" energy, in contrast with "green" or renewable energy.)

We argue that placing caps on large brown energy consumers like data centers can help businesses, utilities, and society deal with these challenges. The caps may be government-mandated, utility-imposed, or voluntary. Governments may impose Kyoto-style *cap-and-trade* schemes to curb carbon emissions and promote green energy. Utilities may impose caps to encourage energy conservation and manage their own costs; in this *cap-and-pay* scenario, consumers that exceed the cap could pay higher brown electricity prices. Finally, businesses may voluntarily set brown energy caps for themselves; in this *cap-as-target* scenario, caps translate into explicit targets for energy conservation. When carbon-neutrality can be used as a marketing tool, businesses may also use caps to predict their expenditures with neutrality and/or green energy.

Regardless of the capping scheme, *the research question is how to create the software support for capping brown energy consumption without excessively increasing costs or degrading performance.* Our work explores this question in the context of Internet services. These services are supported by multiple data centers for high capacity, high availability, and low response times. The data centers sit behind front-end devices that inspect each client request and forward it to one of the data centers according to a *request distribution policy*. Despite their wide-area distribution of requests, services must strive not to violate their SLAs.

We propose and evaluate a software framework for optimization-based request distribution. This framework enables services to manage their energy consumption and costs, while respecting their SLAs. For example, the framework considers the energy cost of processing requests before and after the brown energy cap is exhausted. At the same time, the framework considers the need for high throughput and availability. Furthermore, the framework allows services to exploit data centers that pay different (and perhaps variable) electricity prices, data centers located in different time zones, and data centers that can consume green energy. Importantly, the framework is general enough to be useful even in the absence of brown energy caps, different electricity prices, or green energy.

Based on the framework, we propose request distribution policies for different cap scenarios. Operationally, an optimization-based policy defines the fraction of the clients' requests that should be sent to each data center. The front-ends periodically solve the optimization problem defined by the policy, using mathematical optimization algorithms, time series analysis for load prediction, and statistical performance data from data centers. After fractions are computed, the front-ends abide by them until they are recomputed.

In the remainder of the paper, we briefly describe our framework, request distribution policies, and evaluation results. We refer the interested reader to [1] for an in-depth description of the work.

2. REQUEST DISTRIBUTION POLICIES

2.1 Problem setup. Our goal is to design request-distribution policies that minimize the energy cost of a multi-data-center Internet service, while meeting its SLA. Typically, client requests arriving at the service's front-ends are serviced by a small set (2 or 3) of geographically distributed "mirror" data centers.

For our policies to be practical, it is not enough to minimize energy costs; we must also guarantee high performance and availability, and do it all dynamically. Our policies respect these requirements by having the front-ends (1) prevent data center overloads; and (2) monitor their response times, and adjust the request distribution to correct any performance or availability problems.

We assume that a "power mix" for each center can be contracted with the corresponding power utility for an entire year. The mix defines the percentage of brown and green electricity the utility will pump into the electricty grid. Many utilities allow such arrange-

*This research was partially supported by NSF grants #CSR-0916518 and #CNS-0448070, and Rutgers grant #CCC-235908.

SIGMETRICS'10, June 14–18, 2010, New York, New York, USA.
ACM 978-1-4503-0038-4/10/06.

ments today. The brown energy cap is associated with the entire service (i.e., all of its data centers) and also corresponds to a year. When a service exceeds the cap, it must either purchase carbon offsets corresponding to its excess consumption (cap-and-trade) or start paying higher electricity prices (cap-and-pay). The service has a single SLA with customers specified as (L, P), meaning that at least $P\%$ of the requests must complete in less than L time.

2.2 Optimization-based distribution. We created a framework that includes parameters such as the brown energy cap, the energy cost of serving a request at each data center before and after cap exhaustion, the "base" energy cost of the servers (i.e., the energy cost when servers are idle), and the load capacity of each data center.

We then used the framework to formulate optimization problems (details below) defining the behavior of our request distribution policies for the different cap scenarios. First, our optimization defines the mix of green and brown energy at each data center once per year. Then, at a finer granularity, the optimization determines the fraction of requests to be sent by the front-end devices to each data center during an "epoch". During each epoch, the fractions are fixed. Depending on the solution approach (details below), a set of fractions is computed for one or more epochs at a time. *The computation is performed off the critical path of request distribution.* The front-ends must recompute the fractions if the load intensity, the electricity price, or the carbon offset market prices change significantly since the last computation. Recomputation occurs typically no more than twice per hour.

After each recomputation and/or every hour, the front-ends inform the data centers about their expected loads for the next hour. The data centers use this information to reconfigure, leaving only as many servers active as necessary to service the expected load.

Problem formulations (distribution policies). Based on the framework, we analytically model the overall energy cost of a service under cap-and-trade. The model is based on per-epoch parameters such as the expected number of requests arriving to the service, the percentage of requests forwarded to each data center, the average energy cost of processing a request at each data center, and the base energy cost of each data center under a given load. Our model of energy cost per request includes cap violation charges; that is, beyond the cap exhaustion point, the service has to absorb the additional cost of purchasing carbon offsets on the market. The overall energy cost is minimized under the constraints that the load directed to each data center should not exceed its processing capacity, and the service's SLA must be respected.

We also study a cap-and-pay policy that is similar to its cap-and-trade counterpart, but replaces the purchase of carbon offsets with a fee for cap violation.

Solution approaches to optimization problems. We devised three solution techniques that differ in the extent to which they use predictions and linear programming techniques.

The first technique, Simulated Annealing (SA), first solves the (non-linear) optimization problem to define the power mixes for the year. Then, it solves it for a week at a time to compute the fraction of requests to send to each data center. This approach requires predictions of offered load intensities, electricity prices, carbon market prices, and data center response times for the next week. We use ARIMA modeling for load predictions, and past response time statistics for data center performance prediction.

The next two techniques use linear programming (LP) to solve for the request distribution fractions. They first solve for the power mixes for the year using SA. After that, these approaches compute the fractions only for the next epoch, so that the optimization problem can be converted into an LP problem. The difference between the LP approaches lies in the way they use predictions. The first does not use predictions at all, using the current observed values (e.g., load intensity and response times) as predictors. In contrast, the second approach uses predictions for the next 1-hour epoch.

3. RESULTS

We evaluate our framework and policies through extensive simulations, using a real request load trace from a commercial service and realistic data on energy and carbon offset prices (gathered from various sources on the Web). We have validated our simulator against a real prototype implementation running on servers at four universities (Rutgers, Princeton, Univ. of Washington, and EPFL).

Our results for the cap-and-trade policy demonstrate that our techniques can significantly optimize costs while consistenly satisfying SLAs. In particular, in a scenario with three data centers located at Princeton, UW, and EPFL, and a front-end at Rutgers, our optimization approaches significantly decrease the energy cost (19-35%) compared to a greedy cost-aware heuristic policy. SA consistently outperforms the LP approaches because it optimizes the energy cost considering a much longer horizon (one week vs one hour). In essence, SA can take advantage of periods of low electricity prices at slower data centers, as it predicts that it will be able to compensate later during the week to still meet the SLA.

Our results also demonstrate that our policies allow a service to trade off brown energy consumption and cost. For example, when the brown energy cap is lowered from 100% of the energy needed to process the trace to 75% of that energy, SA can lower brown energy consumption by 24% at only a 10% increase in cost.

Our sensitivity analysis shows that cost and brown energy consumption are most affected by some key parameters, such as the ratio of electricity prices (e.g., brown vs. green and at different data centers), how tight the performance SLAs are (which limits the flexibility of request distribution), and the energy proportionality of data centers.

4. RELATED WORK

To our knowledge, this work is the first to propose capping the brown energy consumption of large computer systems. It is also the first to consider carbon market interactions. Recently, Qureshi studied dynamic request distribution based on hourly electricity prices to reduce the energy cost of Internet services [3]. We introduced a simplified framework to optimize request distributions in [2] in the presence of variable electricity prices and green data centers. However, our previous work did not consider brown energy caps, market interactions, or power mixes; it also did not explore LP-based solution approaches and did not include a real implementation.

5. CONCLUSIONS

We conclude that our optimization framework and policies can play an important role in our increasingly energy-conscious society, as they provide a rigorous approach for services to manage their brown energy consumption and leverage green energy, while respecting their SLAs and minimizing costs.

6. REFERENCES

[1] K. Le et al. Managing the Cost, Energy Consumption, and Carbon Footprint of Internet Services. Technical Report DCS–TR–639, Rutgers University, July 2009.

[2] K. Le et al. Cost- And Energy-Aware Load Distribution Across Data Centers. In *Proceedings of HotPower*, Oct. 2009.

[3] A. Qureshi et al. Cutting the Electric Bill for Internet-Scale Systems. In *SIGCOMM*, August 2009.

[4] US EPA. EPA Report on Server and Data Center Energy Efficiency. August 2007.