Global Cost Diversity Aware Dispatch Algorithm for Heterogeneous Data Centers

Ananth Narayan S.
ans6@sfu.ca

Somsubhra Sharangi
ssa121@sfu.ca

Alexandra Fedorova
fedorova@cs.sfu.ca

Simon Fraser University
Burnaby, Canada

ABSTRACT

Large, Internet based companies service user requests from multiple data centers located across the globe. These data centers often house a heterogeneous computing infrastructure and draw electricity from the local electricity market. Reducing the electricity costs of operating these data centers is a challenging problem, and in this work, we propose a novel solution which exploits both the data center heterogeneity and global electricity market diversity to reduce data center operating cost. We evaluate our solution in our test-bed that simulates a heterogeneous data center, using real-world request workload and real-world electricity prices. We show that our strategies achieve cost and energy saving of atleast 21% over a naïve load balancing scheme that distributes requests evenly across data centers, and outperform existing solutions which either do not exploit the electricity market diversity or do not exploit data center hardware diversity.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems; C.4 [**Computer Systems Organization**]: Performance of Systems; C.5.5 [**Computer Systems Organization**]: Servers

General Terms

Algorithms, Performance

1. INTRODUCTION

The Internet has become ubiquitous, leading to the creation and growth of enormous Internet based companies such as Google, Yahoo, Wikipedia, Facebook, and Amazon among others. In order to fulfill the data needs placed by ever-connected consumers, increasing numbers of data centers are either being setup or leased by these companies across the globe. These data centers are setup in diverse geographic locations with services replicated in the data centers. The location of the data center allows the operator to provide low latency service to customers and the replication supports robustness and reliability. Recently, the electricity cost of operating these data center has emerged as a serious concern for these data center operators; Qureshi et al [10] have recently shown the annual electricity costs for data centers to be in order of several million dollars. Since data center locations are usually geographically far apart from each other, we can expect them to buy electricity from the local markets, where there can be significant variation in electricity price. Moreover, more and more electricity pricing is being done based on hourly consumption basis. Therefore, we have an opportunity to exploit this variation in price by intelligently dispatching the service requests to the less expensive data centers.

While newly setup data centers might be homogeneous in their design, hardware upgrades over time would result in the data center becoming heterogeneous. Heterogeneous computing has been shown to provide improved energy efficiency [4], and we can expect heterogeneity to become increasingly prevalent. This aspect of heterogeneity provides us with an opportunity for cost savings. While operators have control on the heterogeneity of their data center, they do not have control over the electricity prices. Therefore, intelligent solutions that take into account both the heterogeneity of the data center and the energy price diversity are required in the near future.

In this paper we explore techniques to reduce the electricity cost across multiple, heterogeneous data centers. We consider a global multi-location data center service and attempt to minimize the electricity cost at two levels: first, by exploiting the energy price variations in different markets at a global level, and second, by intelligent scheduling for the heterogeneous server hardware within individual data centers. We can leverage this difference in prices to reduce the electricity costs of a data center and provide cost savings in addition to that accrued using energy efficiency measures that each data center may have locally.

First, we consider the problem of load balancing among data centers located across multiple timezones. We also consider the impact of load migration on the quality of service (QoS) for the requests served. While distributing load across data centers can reduce electricity cost by exploiting the lower price in the region from where requests are serviced, it could increase the latency experienced by users resulting in poor user experience. Therefore, it is imperative to keep any QoS loss within tolerable limits. Finally, we investigate the efficacy of a geo-location based strategy. To the best of

ICPE'11, March 14–16, 2011, Karlsruhe, Germany.

our knowledge, this is the first work which considers a combined strategy of exploiting both the energy market price variability and heterogeneity of data center server hardware towards minimizing the electricity cost.

The rest of the paper is organized as follows: A problem description is provided in Section 2 and the proposed solution is explained in Section 3. The methodology is presented in Section 4, experimental results in Section 5, related work in Section 6, and we conclude our discussion in Section 7.

2. PROBLEM DESCRIPTION

We propose a solution to minimize data center electricity cost while at the same time keeping the application response time acceptable by means of intelligently scheduling the incoming service requests.

2.1 Global Data Centers

We denote the electricity cost of a time interval t as $Cost(t)$ and, without loss of generality, assume that the total electricity cost will be minimized if we minimize the cost for each time interval. The electricity cost of a particular time interval is the sum of the costs at each data center. Let there be N data centers and the cost incurred by data center $n \in \{1, \ldots, N\}$ at time interval t be denoted as $Cost_n(t)$. The electricity cost at a data center is a function of the power consumed and the electricity price. Let the unit electricity price at data center n at time t be $\mathcal{E}_n(t)$ and the power consumption during the same interval at the same data center be $W_n(t)$. Hence we have the following optimization function.

$$\text{minimize} \quad Cost(t) = \sum_{n=1}^{N} C_n(t) = \sum_{n=1}^{N} \mathcal{E}_n(t) W_n(t) \qquad (1)$$

The electricity prices for the current interval can be periodically updated at the global load balancer from real time local market predictions known as 'spot price'.

2.2 Data Center Power Consumption

Let us assume that data center n has S_n heterogeneous servers, with H types of servers and an $S^h : h \in H$ number of each type such that $S_n = \sum S^h$. Each server type has a power profile composed of an active power consumption h^{active} and idle power consumption h^{idle}, and a capability profile μ^h; μ^h is the service rate of server type h. Therefore, the total capability of a data center can be given as $N_C = \sum S^h \mu^h$.

When the request rate to a data center equals its service rate, we have full utilization of each server and the power consumed is the sum of active power profiles of all servers. Let us assume that a data center is loaded to x% of its full capacity and let y^h% of that load to be assigned to server type h. Therefore, utilization of server type h can be given as

$$U_n^h(t) = \frac{N_C}{S^h \mu^h} x_n y^h \qquad (2)$$

And the power consumption of each server type h can be given as the sum of active idle power.

$$W_n^h(t) = \left(h^{active} \frac{N_C}{S^h \mu^h} + h^{idle} (1 - \frac{N_C}{S^h \mu^h}) \right) x_n y^h \qquad (3)$$

Symbol	Description
N	Number of data centers
$C_n(t)$	Cost at data center n at time t
$Cost(t)$	Total cost across all data centers
$\mathcal{E}_n(t)$	Unit electricity cost
$W_n(t)$	Power Consumption at data center n
S_n	Number of servers at data center n
H	Type of servers
S^h	Number of servers of type h
N_C	Processing capacity of a data center
h^{active}	Active state power consumption of server type h
h^{idle}	Idle state power consumption of server type h
x_n	Load of data center n
y^h	Load of server type h
μ^h	Processing capacity of server type h
μ_n	Average processing capacity of data center n
D_n	Delay incurred by data center n

Table 1: List of symbols used in the paper

Summing over all server types, the total power consumption of the data center n is obtained as

$$W_n(t) = \sum_{h \in H} \left(h^{active} \frac{N_C}{S^h \mu^h} + h^{idle} (1 - \frac{N_C}{S^h \mu^h}) \right) x_n y^h \qquad (4)$$

For brevity, we drop the time suffix for the rest of our discussion.

2.3 Application Response Time

The important quality metric of a dispatching algorithm is the average delay for servicing a request which is defined as sum of the waiting time in the server queue and the processing time. In a homogeneous server setup, all servers are assumed to have equal, high service rate, and the processing time component can be ignored. However, in a heterogeneous setup the processing time can be a significant component, specially for the slower machines. While there are several models present for analyzing homogeneous server systems, modeling of heterogeneous server systems is complex and we approximate the heterogeneous servers using a homogeneous model so that existing results can be applied. We take the average service rate of the heterogeneous servers to be the service rate of the data center. Let μ_n be the average service rate of the heterogeneous servers for a data center load of x% and y^h% load for server type h. We have:

$$\mu_n = \sum_{h \in H} \mu^h x_n y^h. \qquad (5)$$

We already know the request arrival rate for each data center $\lambda_n = N_C x_n$. Therefore, from queuing theory results [13], the average time that a request stayed in the system can be given as

$$D_n = \frac{1}{\mu_n - \lambda_n} = \frac{1}{\sum_{h \in H} \mu^h x_n y^h - N_C x_n} \qquad (6)$$

Equation 6 suggests that an increase in the request rate will result in an increase in the time spent by each request in the system, which is intuitively correct. We note that, for a work conserving system(i.e. where no jobs are dropped after they are allocated to servers) the allocation of requests to a server can never exceed the servers processing capacity. We

enforce this as the following constraint:

$$N_C x_n y^h < \mu^h \quad (7)$$

2.4 Load Distribution Constraints

So far, we have encountered two unknowns in our model.

1. The global load balance factor x (calculated at the front end) which determines how many requests are to be routed to which data center.
2. The server load balance factor y which is determined locally at each data center.

Let the request rate seen at the front end be λ. If there are N data centers with equal capacities N_C, and they are loaded to x_n% of their capacity by the dispatching algorithm, then the load at each data center is $\lambda_n = N_C/x_n$. The individual loads must add up to the total incoming load (seen at the front end) as the following.

$$\lambda = \sum_{n=1}^{N} \lambda_n = \sum_{n=1}^{N} N_C x_n \quad (8)$$

Inside each data center we have H server pools, each serving y^h% of the incoming load. Therefore, the requests served by each server pool must add up to the total incoming load for that data center. Hence we have,

$$N_C x_n = \sum_{h \in H} y^h \quad (9)$$

3. PROPOSED STRATEGIES

3.1 Global Dispatch Strategies

In this work, we evaluate the following global load balancing strategies : Even Distribution, Least Electricity Price, Latency Aware, and Combined distribution. Each of these strategies is described below.

3.1.1 Even Distribution

Even Distribution is a naïve strategy where the requests are distributed evenly across all available data centers; timezone, electricity costs, or latency are *not* considered. Consequently, we can expect this strategy to perform poorly when compared to other strategies discussed below. This strategy provides us the baseline to compare other strategies that consider the electricity price.

3.1.2 Least Electricity Price

In the Least Electricity Price distribution, requests are redirected to the data center located in the region where the current price of electricity is the least across all timezones and data centers. With such a strategy, we can clearly reduce the electricity costs associated with servicing the requests. We name the data center at the location which currently has the lower electricity price as the 'primary data center'. Requests are dispatched to data centers in the increasing order of electricity price if the primary data center is operating at full capacity.

3.1.3 Latency Aware

In the latency aware distibution strategy, the source of the requests is considered and requests originating from a geographical region are serviced by the data center that exhibits the lowest network latency to that region. Intuitively, one can expect that requests are serviced by the data center that is geographically closest to their point of origin. However, differences in bandwidth can impact the latency, resulting in requests being serviced at a data center located farther away from the geographically nearest data center.

3.1.4 Combined

The hourly electricity prices in each data center location and the network latency for the request are both considered in the scheduling decision in this strategy. Consequently, we can expect the combined global load balancing strategy to perform the best in terms of both cost reduction and latency reduction compared to the other global scheduling strategies introduced earlier.

3.2 Local Scheduling

Further to the global scheduling, where requests are distributed between global data centers, we evaluate two 'local' scheduling strategies - strategies used within a data center to schedule requests to servers.

3.2.1 Proportional Fair (PF)

In this strategy, the requests are divided across all servers in proportion of their service rates, where the service rate captures the maximum number of requests that a server can handle in unit time. Different server types execute different number of requests, however, all servers in the data center exhibit comparable utilization; utilization, in this case, being the ratio of the requests executed to the server's service rate. In our setup, we use three types of machines - Intel Core i7 (Nehalem), Intel Atom, and ARM BeagleBoard (Section 4.2).

3.2.2 Fastest Server First (FSF)

In this strategy, requests are assigned to the fastest servers before they are assigned to the slower servers. With the FSF strategy, requests are sent to Nehalem machines which provide a higher service rate. The requests that cannot be assigned to the Nehalem machines are then assigned to the Atom machines, and if the requests exceed the service capacity of both Nehalems and Atoms, then they are assigned to the BeagleBoards.

3.3 Local Power Management

Thus far we have discussed the global load balancing and local scheduling strategies. In addition to the scheduling, we simulate a simple power management algorithm at each data center. The power management algorithm calculates the number of servers required to service the current requests and the rest of the machines are deactivated. The inactive servers do not consume any power and consequently the cumulative power consumption of the data center is reduced. For the Proportional Fair local scheduling, an equal number of servers of each type are active, whereas for the FSF scheduling, the lower capacity servers are inactive unless the requests cannot be serviced by the higher capacity servers.

4. EVALUATION METHODOLOGY

Our evaluation involved simulation of the proposed scheduling strategies using data centers located in two different timezones, a real world workload trace, and real world electricity prices.

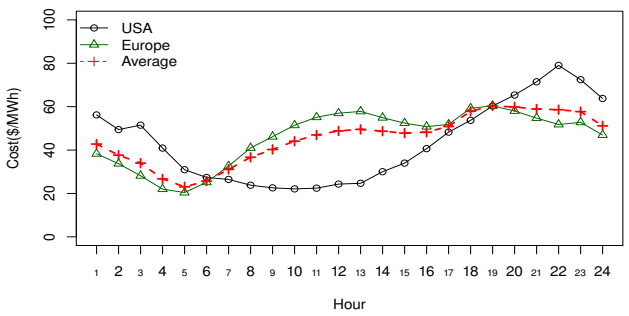


Figure 1: Time shifted, hourly electricity prices

4.1 Workload

We use the request trace from Wikipedia [12] to simulate our workload. For our evaluation, we require the daily variation in the number of requests seen over time and the percentage of total requests received from different countries. Wikipedia statistics site[3] provides statistics on the number of requests received from each country and the percentage of the total load that the requests constitute. From the entire list, we selected the subset of countries which had a contribution of at least 50 million requests and aggregated these countries to 5 regions - North America, Latin America, Europe, Asia-Pacific, and Asia. The requests originating from the selected countries accounted for 89% of the total requests handled by Wikipedia (Table 2).

4.2 Data Center Setup

We consider a heterogeneous setup, with each data center consisting of three types of servers - Intel Core i7 (Nehalem), Intel Atom, and ARM BeagleBoard - and assume an equal number of machines of each type in each data center. Each of the server types exhibits a different service rate and energy consumption footprint [6]. Requests typically show a diurnal pattern, where the load is high in some parts of the day and low during others. Consequently, we can expect to see variation in the utilization of the servers over time. We assume a linear relationship between utilization and active power consumption; idle power consumption makes a fixed contribution (Equation 3). Wikipedia uses two hosting facilities in Florida, USA and two in The Netherlands; in our setup, we use the same timezones but assume one data center in each timezone, and also assume that a front-end device balances the incoming traffic between the hosting facilities.

4.3 Hourly Electricity Prices

We obtained per-hour electricity prices from The USA [1] and The Netherlands [2] for an entire week. Analyzing the data, we observed that across the seven days, the electricity price for the same hour of the day can show substantial fluctuation. Therefore, we averaged the hourly prices for the seven days for our calculations.

Figure 1 shows the hourly variation in electricity prices, measured in Dollars per Mega Watt hour ($/MWh). The prices are corrected for timezone differences. The line representing Europe (marked with o) shows the variation in Europe region from 00:00 to 23:59 Central European Summer Time (CEST) for a Wednesday. Prices from USA Eastern Standard Time (EST) (marked with △) are timezone corrected to use 6 hours from Tuesday evening (18:00 to 23:59) and 18 hours of Wednesday(00:00 to 17:59; the line marked with '+' plots the average price. We can see that prices in Europe are higher for 13 hours and the prices in the USA are higher for the remaining 11 hours in a span of 24 hours.

4.4 Network Delay Estimation

We obtained the ping latency from each of the 5 regions listed in Section 4.1 to two target regions - North America and Europe. Table 2 captures the ping latency (measured in milli seconds) from each of the source regions to two destination regions North America and Europe.

Source	North America	Europe	% Requests
North America	41	95	37
Europe	95	16	32.35
Asia Pacific	109.6	125.75	13.7
Latin America	124.75	150.75	3.80
Asia	273.0	125.6	1.9

Table 2: Network Latency and Request Distribution

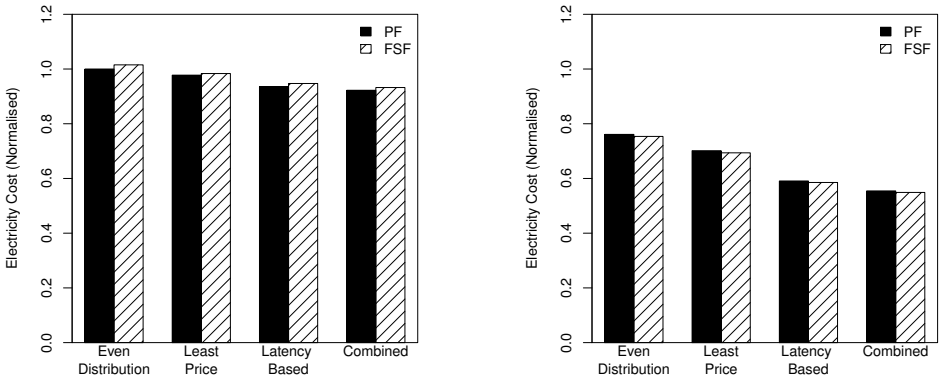
5. EXPERIMENTAL RESULTS

In this section we present and discuss the results of our simulations. We investigate the total electricity cost, power consumption, and the latency resulting from the strategies. Based on the discussion of the strategies presented earlier, we expect the Even Distribution strategy (which naïvely assigns requests equally between available data centers) to show the highest electricity cost of all four strategies evaluated. We normalise all results to the Even Distribution strategy (which is also our baseline global strategy), with the Proportional Fair local scheduling algorithm and no local power management.

Figure 2 and Figure 3 capture the normalised total electricity cost and total power consumption incurred to complete the trace, i.e., for 24 hours of operation of two data centers respectively. Each figure has two sub figures - sub figure (a) captures the results when *no power management* was simulated, and sub figure (b) captures the results with *power management active* at each data center. Under each global load balancing strategy, the FSF local strategy is shown in white, and PF is shown in black.

As mentioned earlier, we expect the Even Distribution load balancing strategy to perform poorer than the other strategies because it does not take either price or latency into decision making and naïvely distributes the load evenly across the two data centers. We observe in Figure 2, that the total electricity costs for Even Distribution is never lower than any of the other strategies. The Combined load balancing approach, which considers both electricity price and network latency, exhibits the least cost. The power consumption (captured in Figure 3) of four global load balancing strategies are comparable, with the Combined scheduling strategy exhibiting the least power consumption. When data centers do not have any local power management, the power consumption of the idle servers is accounted for, resulting in the comparable behaviour.

With the presence of local power management, the load balancing strategies show a marked difference in costs resulting from the reduction in power consumption - from a min-



(a) No Power Management

(b) With Power Management

Figure 2: Electricity Cost.

imum of 21% (Least Price, PF) to a maximum of 38% (Latency Aware, FSF) reduction in power consumption (Figure 3b). Analysing the performance of the four strategies when executed with local power management, we observe that the Combined strategy outperforms others, exhibiting the least electricity price and power consumption - 23.75% lower electricity costs (Figure 2b)and 26.3% lower power consumption (Figure 3b)compared to the Even Distribution, Proportional Fair local scheduling.

All regions barring Asia and Europe, show a lower latency to reach North America (compared to Europe); requests from these regions also constitute 55% of the total. North America also shows lower electricity price for 13 hours. These factors together result in the Latency Aware strategy performing comparably with the Combined strategy because for 13 hours, the data center that exhibits low latency also has the lower electricity price.

Between the local scheduling strategies, the PF local scheduling policy shows between 1 and 2% reduction in power and electricity cost compared to the FSF strategy when no local power management in present. However, with power management active, PF exhibits 1% increase in power compared to FSF. In the presence of power management, keeping the 'faster' servers busy is more energy efficient than distributing the load across all server types.

Figure 4 captures the latency (measured using the ping latency discussed earlier) experienced by users under each scheduling strategy. This graph does not capture the processing delay or the time spent by the request in the data center waiting for service and being serviced. For simplicity of visualization, we divide the latency values into bins of 50 ms each, except for the last bin which captures latency above 300 ms. With the Latency Aware and Combined scheduling strategies 75% of the requests see a latency of under 50 ms. The remaining 25% requests experience a latency of 100 ms or higher. The other two strategies – Least Price and Even Distribution – show comparable counts in each latency bin; about 38% requests experience under 50 ms latency and an equal number experience between 51 and 100 ms latency. The requests originating from the North America and Europe regions constitute two-thirds of the requests received in the entire trace. The data centers in our simulation are also

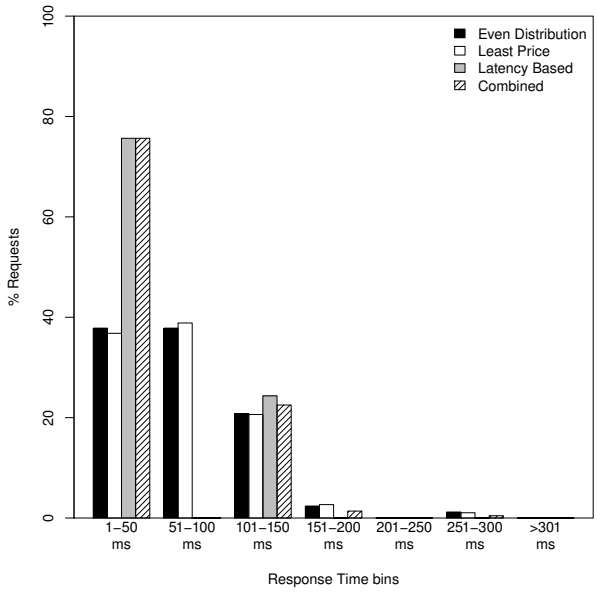
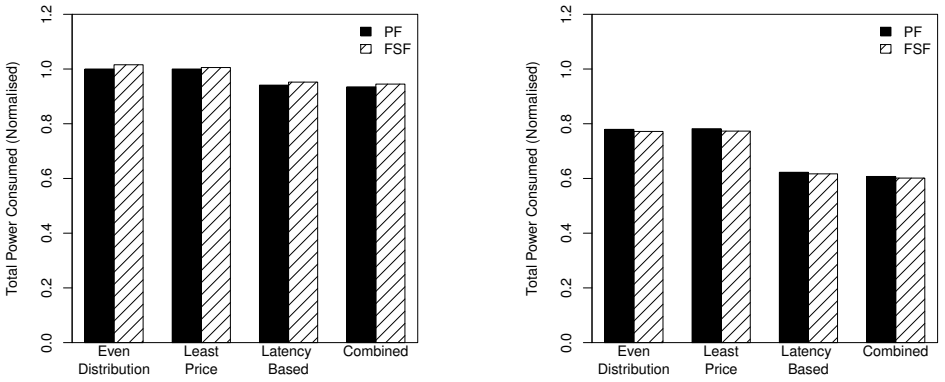


Figure 4: User Latency

located in the same regions, and the requests stay within the region, thus showing low latency.

6. RELATED WORK

Reducing power consumption costs in data centers has primarily focused on power managing servers using processor dynamic voltage/frequency scaling (DVFS) and placing servers in sleep states [5] while another dimension in reducing power has been in reducing the cooling costs associated with operating a data center [8]. However, leveraging electricity prices is a new aspect affecting data center operational costs. In our work, we attempt to leverage differences in electricity prices across timezones to minimize the electricity costs of data centers. The work that come closest to ours are the by Qureshi et al [9], Le at al [7], and Rao et al [11].



(a) No Power Management

(b) With Power Management

Figure 3: Power Consumption

Qureshi et al [9] investigated the cost savings accrued by leveraging the temporal and geographical variation in electricity costs and shifting computation across energy markets. While this study makes a case for leveraging electricity prices and reducing costs, it does not investigate the performance impact of load shifting. Le et al [7] investigated load balancing across global data centers utilizing green and and brown energy costs. They apply their solution over large time windows - in the order of hours; our approach attempts to use shorter time windows. Rao et al [11] do not consider the differences in timezones and also assume that data centers exhibit constant power consumption.

Prior work on global load balancing has focused on homogeneous data centers, where all machines in the data center exhibit the same power and performance profiles; our work is based on heterogeneous data center setup consisting of three different types of machines. Further, we do not assume that the servers are always busy, but vary in their utilization over time - due to the diurnal variation in load, and based on the scheduling algorithm used in data center, which is representative of real world conditions.

7. CONCLUSION

In this work, we investigated different global load balancing strategies and compared them on power consumption and electricity cost incurred to execute a 24-hour workload trace. With each global load balancing strategy, we investigated two local scheduling strategy - Proportional Fair (PF) and Fastest Server First (FSF). The Combined global scheduling strategy performed the best, exhibiting lower power consumption and consequently lesser electricity cost, and also exhibited the least latency. Additional aspects that can be introduced into the model are communication costs and the cost of electricity generation that captures the carbon footprint of each electricity source.

8. REFERENCES

[1] Ameren. `www.ameren.com`.

[2] APX-Endex. `www.apxendex.com`.

[3] Wikipedia Statistics Website. `stats.wikimedia.org`.

[4] Barroso, L. A., and Hölzle, U. The case for energy-proportional computing. *Computer 40*, 12 (2007), 33–37.

[5] Bianchini, R., and Rajamony, R. Power and energy management for server systems. *Computer 37*, 11 (2004), 68–74.

[6] Krioukov, A., Mohan, P., Alspaugh, S., Keys, L., Culler, D., and Katz, R. NapSAC: Design and Implementation of a Power-Proportional Web Cluster. In *Proceedings of the First ACM SIGCOMM Workshop on Green Networking* (Aug 2010).

[7] Le, K., Bilgir, O., Bianchini, R., Martonosi, M., and Nguyen, T. D. Managing the cost, energy consumption, and carbon footprint of internet services. *SIGMETRICS Perform. Eval. Rev. 38*, 1 (2010), 357–358.

[8] Moore, J., Chase, J., Ranganathan, P., and Sharma, R. Making scheduling "cool": temperature-aware workload placement in data centers. In *ATC '05: Proceedings of the annual conference on USENIX Annual Technical Conference* (2005), USENIX Association.

[9] Qureshi, A. Plugging Into Energy Market Diversity. In *7th ACM Workshop on Hot Topics in Networks (HotNets)* (Calgary, Canada, October 2008).

[10] Qureshi, A., Weber, R., Balaskrishan, H., Guttag, J., and Maggs, B. Cutting the Electric Bill for Internet-Scale Systems. In *ACM SIGCOMM* (Barcelona, Spain, August 2009).

[11] Rao, L., Liu, X., Xie, L., and Liu, W. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *INFOCOM, 2010 Proceedings IEEE* (14-19 2010), pp. 1 –9.

[12] Urdaneta, G., Pierre, G., and van Steen, M. Wikipedia workload analysis for decentralized hosting. *Comput. Netw. 53*, 11 (2009), 1830–1845.

[13] Yu, S., Doss, R., Thapngam, T., and Qian, D. A transformation model for heterogeneous servers. *High Performance Computing and Communications, 10th IEEE International Conference on* (2008), 665–671.