Cooling-aware Geographical Load Balancing Visualized*

Michael Hirshleifer
California Institute of Technology
1200 E California Blvd
Pasadena, California
111mth@caltech.edu

Yizhen Wang
California Institute of Technology
1200 E California Blvd
Pasadena, California
ywang3@caltech.edu

Zhenhua Liu
California Institute of Technology
1200 E California Blvd
Pasadena, California
zliu2@caltech.edu

Adam Wierman
California Institute of Technology
1200 E California Blvd
Pasadena, California
adamw@caltech.edu

ABSTRACT

We explore how geographical load balancing can improve the efficiency of renewable energy use in data centers. The model incorporates varying cooling efficiency (considering weather conditions) and electricity prices over time at each data center. We run a convex optimization over routing, using as input real workload, temperature, solar, and wind traces. We find that using geographical load balancing lets data centers more effectively use locally available renewable energy, thereby substantially reducing their usage of grid electricity. This conclusion holds across seasons. We develop a visualization that displays the demand from each of the 48 contiguous U.S. states and the energy usage, grid energy usage, and renewable energy generation at each of 10 data centers, animated over time according to the input data and the optimization output. The resulting software can be used to test effectiveness of and refine routing algorithms.

1. INTRODUCTION

1.1 Data centers

Internet companies such as Google use data centers (DCs) to service users' requests (such as Google searches). Each data center contains the computer servers (including processors to perform calculations and disks to store and retrieve user data) that service the requests, along with the infrastructure needed to support their operation (power supply, cooling, networking). Google maintains on the order of one million servers [3] [12] across at least 17 data centers [6].

*We are grateful to Professor Adam Wierman for his insightful and patient guidance. This research is supported by the NSF, Mr. David C. Elliot's family, and the Rose Hills Foundation.

1.2 The problem: Energy cost is important for DCs

Servers use energy to process requests. Data centers require a substantial amount of energy, both for processing requests and for cooling. The energy the servers themselves use to process requests varies with load (requests processed per second).

In 2010, data centers' power usage was between 23 and 31 gigawatts[1], constituting approximately 1.3% of world electricity consumption, and 2% in the US. Their consumption is growing rapidly, having approximately doubled from 2000 to 2005, and increased by 30 to 50 percent through 2010. Google used approximately 220 megawatts during 2010. [12]

Computing companies purchase much of this electricity off the grid. As energy usage is often the largest contributor to cost for running a data center, reducing it would save large amounts of money. Further, the electrical grid provides mainly "brown" energy, which causes pollution; DCs were estimated to have contributed 0.5% of U.S. greenhouse gas emissions in 2007 [9].

Cooling also uses energy. Data centers also spend a considerable amount of energy on cooling [10], especially in summer. The energy needed for cooling largely depends on the weather at data center locations.

Building cost-saving and environmentally-friendly data centers has become a pressing challenge for the ICT industry. Energy consumption accounts for much of a data center's running cost [10], and the carbon dioxide emissions consequent to data centers' grid energy usage concerns society at large. As the need for data centers grows rapidly, "green" routing algorithms would benefit internet companies by reducing their electricity costs, and everyone else by reducing pollution from non-renewable energy. The social benefit

[1]Other estimates: 38 GW in 2012 [7]; "Total power consumption of around 31 GW.[...]A projected rate of increase in energy consumption of 19% into 2012" [4]

could be substantial, as data centers are expected to use several percent of U.S. and world electricity output, and they currently use mostly non-renewable energy.

1.2.1 Load balancing between DCs

There are often several redundant data centers, each of which can process a user request, such as a Google search. Traditionally, a request would be routed to the geographically-nearest data center, thereby minimizing latency—without considering energy efficiency.

1.2.2 Load, energy availability, grid electricity prices, etc. are time-varying and stochastic.

Renewable vs. grid energy. Currently, data centers are powered mainly by buying electricity off the grid. Grid electricity necessarily comes substantially from burning fossil fuels, because with current technology, storing renewable energy over time to match demand is prohibitively expensive.

However, some data centers are already partly operated on green energy. Data centers may have dedicated renewable energy generation (solar or wind farms). But the renewable power output at each location varies over time (time of day, cloud cover, wind speed), and generally does not match user demand, so excess demand must be satisfied by buying electricity off the grid.

1.3 Previous work

1.3.1 Geographical load balancing

Flexibility in routing could allow a network of data centers to run almost entirely off renewable energy. Particularly, previous research suggests that under the geographical load balancing (GLB) model, this can be done without undue increase in latency [13].

When routing a request, there is a trade-off: instead of always choosing the nearest data center, a request can be routed to a data center where renewable energy is currently being produced (reducing electricity costs), at the expense of somewhat higher latency. Professor Wierman's group have devised optimal or near-optimal (given a functional form for the business cost of increased latency) routing algorithms for efficient geographical load balancing.

Convex optimization framework; solve numerically. Our simulation models the whole data center routing system, using real-world data on wind and solar power output, grid electricity prices, and internet request volume over time at various locations (the 48 contiguous U.S. states and 10 data center locations). The simulation is implemented as a convex optimization problem, solved numerically, whose solution is an allocation of requests to data centers.

1.4 Visualization

We implement a comprehensible visualization of the multi-dimensional simulation input and output data. The visualization makes it easier to see what each optimization algorithm is doing, so it can be used to test effectiveness of and refine routing algorithms. This will help with developing an implementable algorithm / system for routing internet requests that allows data centers to operate using almost exclusively renewable energy.

1.5 Cooling-aware GLB

The previous GLB model ([13]) only considers the energy cost of running the servers themselves. But in reality, a data center expends a considerable amount of energy to dissipate the heat the servers produce.

The GLB concept also applies to cooling. That is, we can also exploit the geographical heterogeneity of the data centers to reduce cooling cost, by routing requests to where cooling is currently cheap. We provide a new geographical load balancing model which considers the energy cost of cooling, and show that it can reduce the total cost significantly compared to previous models. This improvement in green energy usage leads to much lower carbon emissions.

We set up an experiment to investigate the economic and environmental impact of using our new model. We use real traces of data center workload, renewable energy availability, and weather to numerically compute the optimal solution.

Our study leads to three major findings.

First, our new GLB model, the Cooling-aware GLB model, reduces the total cost of the system. This is because it is able to distribute the request to locations with not only cheap energy, but also favorable weather for cooling. More excitingly, the optimal cost does not vary much with seasonality. This can potentially solve the problem of data center cooling in places with climate and weather extremes (like strong diurnal temperature variation).

Second, the Cooling-aware GLB model can allow significantly reduced carbon emissions, halved compared to the previous models. It does this by making better-informed decisions than the previous GLB model by considering the energy for cooling.

Third, considering cooling yields a larger benefit than adding energy storage, both in total cost and in carbon emissions, provided that the aggregate renewable supply is enough to power the data centers, yet not in vast surplus. We obtain this result by systematically comparing the Cooling-aware GLB system with storage systems of various capacities. We conclude that Cooling-aware GLB has a clear advantage over the storage model when the renewable energy generation and storage facilities are not widely established; it requires much less investment in infrastructure and hence can be implemented more easily.

2. SETUP

We assume that each data center has a cooling system described in [15], and then modify the model in [13] to include energy cost of cooling. The model can be solved using a convex optimization technique, as in [14].

2.1 The workload

Let J be the set of sources of requests. Each of 48 U.S. states is modeled as having a source of web requests located

at its geographical center. The request volume is $L_j(t)$, the mean arrival rate from source $j \in J$ at time interval t. $L_j(t)$ is estimated using real-world traces. The workload for each source is based on a trace at Hewlett-Packard Labs, scaled by the population of the state, and temporally shifted according to time zone.

2.2 The availability of renewable energy

There are three major considerations when modeling the availability of renewable energy:

First, the weather at the data centers is crucial, as it determines the renewable energy output. We use real traces of wind speed and insolation obtained from [1] and [2]; power output at each data center is proportional to the corresponding weather variable. The measurements have a granularity of 10 minutes.

Second, the size of the renewable energy plant matters. We scale the renewable generation capacity such that the total renewable energy is c times as much as the total energy demand to process all requests. In our simulation, c varies in $[0.5, 4]$ with a step length of 0.5.

Third, we must choose an optimal mix of solar and wind energy to power the data centers. Previous research [13] suggests that a mix of 80% of wind and 20% of solar fits the data center energy consumption characteristics nicely. We adopt this ratio in our simulation throughout.

2.3 The internet-scale system

The internet-scale system consists of a set $\mathcal{N}$ of 10 data centers at Google data center locations in California, Washington, Oregon, Illinois, Georgia, Virginia, Texas, Florida, North Carolina, and South Carolina. The number of servers (or equivalently, processing capacity) X_i at data center i is set to be twice the peak load at i when all requests are routed to the nearest data centers.

The optimization will determine the routing plan $\lambda_{ij}(t)$ and the number of active servers $x_i(t)$ that minimizes the total cost (the sum of delay cost, energy cost, and switching cost).

2.3.1 Delay cost

The delay cost represents the business revenue loss incurred due to delay in processing requests. It comprises the propagation delay d_{ij} from source j to date center i and the queuing delay at i. The propagation delay d_{ij} is calculated to be the time needed to travel between i and j at a transmission speed of $200\frac{\text{km}}{\text{ms}}$ plus a constant term 5ms. The queuing delay is calculated from the parallel M/G/1/Processor Sharing queue model in which the total load $\lambda_i(t) = \sum_j \lambda_{ij}(t)$ is distributed evenly across $x_i(t)$ homogeneous servers of service rate $\mu_i = 0.1/\text{ms}$.

2.3.2 Cooling optimization

The cooling optimization model finds the minimum energy consumption required to maintain the data center at constant temperature $T = 25°\text{C}$. A typical data center uses both air cooling and chilled-water cooling. Let $x = x_a + x_c$ be the total number of active servers, x_a the number of those that are air-cooled, and x_c the number cooled with chilled water. This model finds the best division between x_a and x_c.

The energy consumption of air cooling is

$$c_a(x_a) = k \cdot {x_a}^3, 0 \le x_a \le \bar{x}, k > 0 \tag{1}$$

The parameter k is proportional to the temperature gradient between the inside and outside air. The parameter $\bar{x}$ corresponds to the maximum number of the servers that can be cooled by air cooling alone. The cap $\bar{x}$ is proportional to both the temperature gradient and the maximum air flow rate. In our simulation, the air flow rate is set such that when the outside temperature is 20°C lower than T, the data center can rely on air cooling entirely at full workload.

On the other hand, the energy consumption of chilled-water cooling is empirically roughly linear in the IT demand, so we model it as

$$c_c(x_c) = \gamma \cdot x_c \tag{2}$$

For notational convenience, define

$$u^+ \stackrel{\text{def}}{=} \max(0, u) \tag{3}$$

to be u clamped to be nonnegative.

The optimal cooling portfolio is

$$c(x) = \min_{x_a \in [0,x]} \left(\gamma \cdot (x - x_a)^+ + k \cdot {x_a}^3\right) \tag{4}$$

which yields

$$c(x) = \begin{cases} k \cdot x^3 & \text{if } x \ge x_s \\ k \cdot x_s^3 + \gamma \cdot (x - x_s) & \text{otherwise} \end{cases}$$

where $x_s = \min\left(\sqrt{\gamma/3k}, \bar{x}\right)$ is the threshold for when chiller cooling is necessary.

To explore the effect of seasonality on the performance of the cooling model, we use two sets of temperature data from [8]. One is taken from the first week of January 2012; the other is taken from the first week of July 2012. The measurements are taken hourly.

2.3.3 Energy cost

The energy cost is the cost of both running active servers and keeping them at constant working temperature. The data centers pay no cost for renewable energy; we assume that each data center has dedicated renewable energy generation facilities that incurs zero maintenance cost. Thus the energy cost is for using energy from the grid and can be represented as

$$p_i \cdot (l(x_i(t)) + c(x_i(t)) - r_i(t))^+ \tag{5}$$

where p_i is the price of electricity; $x_i(t)$ is the number of active servers at that time interval t; $l(x_i(t))$ is the energy consumption of active servers themselves in the time interval, or IT demand; $c(x_i(t))$ is the energy usage for cooling; and $r_i(t)$ is the renewable energy availability. In our model, p_i is set to be constant according to the real statistics of each state; l is a linear function of $x(t)$.

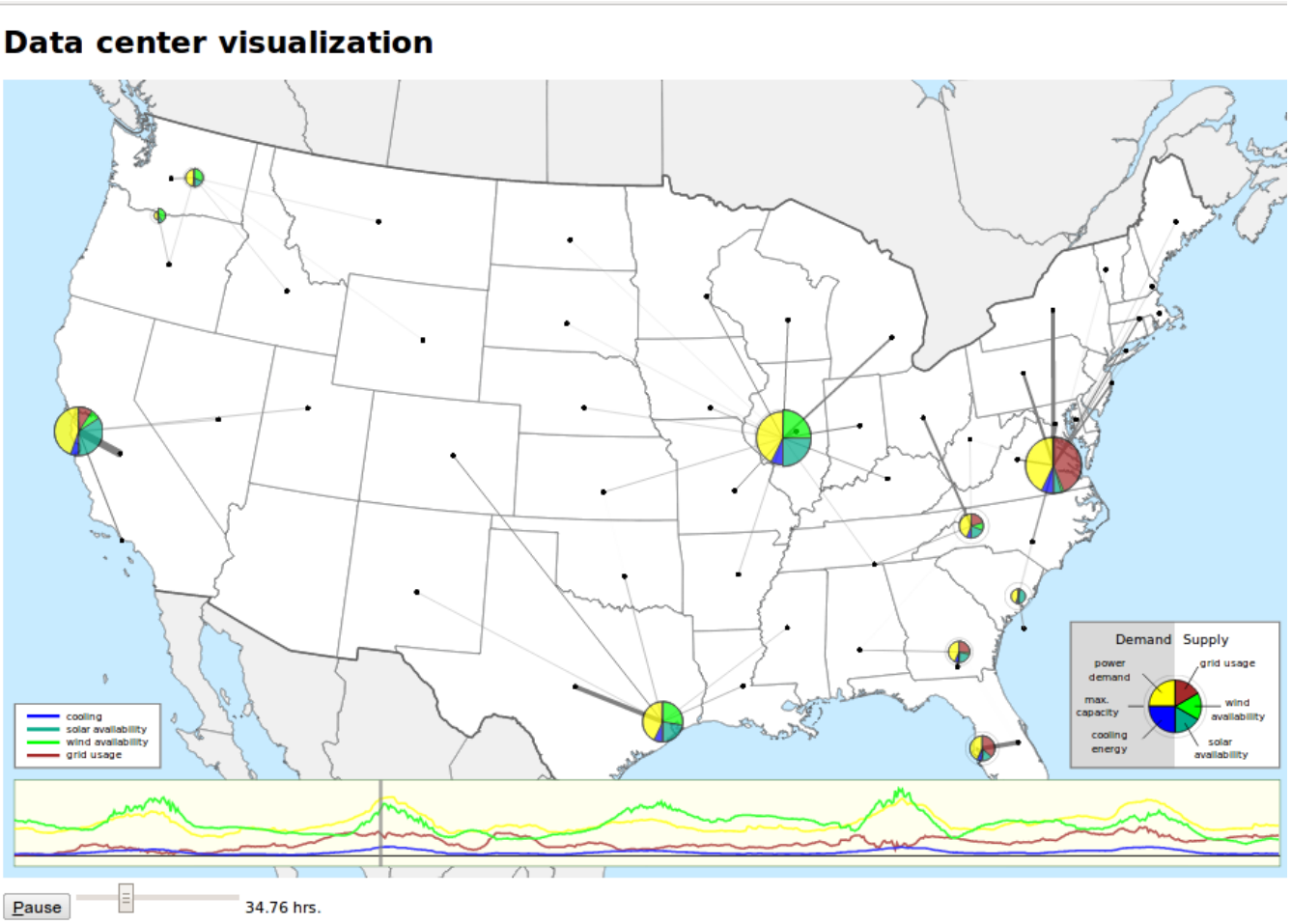


Figure 1: Screenshot of the visualization, running in the Chromium web browser.
The interactive visualization can be viewed at `http://geographicalloadbalancing.github.com/svgmap.xhtml`. The user may play or pause the animation using the button, and seek through the animation using the slider, at the bottom of the page.

2.3.4 Switching cost

The switching cost models the delay and wear-and-tear cost when switching on/off servers. In our model, the workload at each data center is updated every 10 minutes. To avoid excessively-frequent switching of server status, we define the switching cost to be

$$\beta \cdot (x_i(t+1) - x_i(t))^+$$

where β is the weight of switching cost. In our simulation, $\beta = 6$.

2.3.5 Storage

In addition to renewable-energy generators, data centers may also install energy storage (in the form of batteries, supercapacitors, flywheels, etc.). Renewable-energy availability varies greatly over time; introducing storage smooths out the supply.

We model the amount of electricity storage at time t to be $0 \leq es_i(t) \leq ES_i$, where ES_i is the maximum storage capacity. Let the storage rate be

$$e_i(t) = \rho \cdot (es_i(t) - es_i(t+1))$$

Positive $e_i(t)$ means discharging, negative charging. The parameter ρ represents the charging and discharging efficiency. In our simulation we assume perfect efficiency, i.e. $\rho = 1$.

In the presence of storage, the energy cost is

$$p_i \cdot (l(x_i(t)) + c(x_i(t)) - r_i(t) - e_i(t))^+ \tag{6}$$

2.3.6 Total cost

Now we can formally write our optimization problem as:

$$\begin{aligned}\min_{\mathbf{x(t)},\lambda(\mathbf{t})} & \sum_{i\in\mathcal{N}} p_i \cdot (l(x_i(t)) + c(x_i(t)) - r_i(t) - e_i(t))^+ \\ & + \sum_{j\in\mathcal{J}} \sum_{i\in\mathcal{N}} \lambda_{ij}(t) \left(\frac{1}{\mu_i - \lambda_i(t)/x_i(t)} + d_{ij} \right) \\ & + \beta \cdot (x_i(t+1) - x_i(t))^+\end{aligned}$$

subject to

$$\begin{aligned}
&\sum_{i\in\mathcal{N}} \lambda_{ij}(t) = L_j(t), && \forall j \in J \\
&\lambda_{ij} \geq 0, && \forall i \in N, j \in J \\
&0 \leq x_i(t) \leq X_i, && \forall i \in N \\
&\lambda_i(t) \leq x_i(t) \cdot \mu_i && \forall i \in N \\
&0 \leq es_i(t) \leq ES_i && \forall i \in N \\
&e_i(t) = es_i(t) - es_i(t+1) && \forall i \in N
\end{aligned}$$

The boundary conditions model the following real-world constraints:

1. The requests from a population center all have to be processed by the geographical load balancing system. Each data center should receive a non-negative amount of work.
2. Each data center has a limited number of servers; it cannot process requests in excess of its computational capacity.
3. Each data center's energy storage capacity is bounded between 0 and a fixed cap.

2.3.7 CO_2 emissions

The optimization thus far has the sole objective of minimizing the cost to the firm. But we also want to monitor the amount of CO_2 emitted as an externality. The emission of CO_2 arises from grid electricity use; the emissions rate per kW·h varies across states, as each state has a different energy-source composition. An estimation of these rates can be found at [5].

Letting η_i be the rate of carbon emission per kW/h at data center i, the total CO_2 emission is then

$$\sum_{i\in\mathcal{N}} \eta_i \cdot (x_i(t) + c(x_i(t)) - r_i(t) - e_i(t))^+$$

2.3.8 Benchmarks for comparison

Our model is the first to integrate practical cooling concerns into geographical load balancing. To show their impact, we will compare the result of our model (Cooling-aware GLB) to the previous Cooling-oblivious GLB model and the model which simply routes all requests to the nearest data centers (LOCAL). For the two benchmark models, the cooling optimization is done after the routing scheme is determined.

Separately, we compare the performance of our model to the storage model. Storage capacity is quantified as the duration that a data center could operate at maximum load entirely off stored energy (excluding energy usage for cooling). We assume that storage incurs no running cost. We choose data-center systems with 3-hour and 6-hour storage respectively as the benchmark.

We will evaluate the outcomes based on 1) the total cost; 2) the grid (brown) energy usage; and 3) the CO_2 emissions.

3. VISUALIZATION

Our numerical experiment generates time series describing routing plans and data center activity. Due to the large quantity of data, the output is hard to interpret when in matrix form. Therefore we developed a visualization tool to help us recognize the key characteristics of the optimal solution quickly. It also works as a simple check for the validity of the solution: we can spot obviously abnormal behaviors in the output.

3.1 Technical description

The output visualization is in Scalable Vector Graphics (SVG) format, a high-quality XML-based vector graphics format supporting fluid animations that display time-series data. This animation can be embedded into a web page.

We also wrote a wrapper XHTML web page to make the simulation easier to use. The user can control the animation (play, pause, or seek into the animation) using the interface on the web page (implemented using Javascript). Possible enhancements could include zooming, panning, and showing/hiding types of map elements.

Both components work in the Chromium web browser and pass the W3C validators.

The back-end script that generates the SVG map from input data is written in the Scala programming language. The script reads in and displays data center locations, client (request source) locations, real solar- and wind-generation traces, and the routings optimized according to the various algorithms (from the output of the Matlab optimizations). We are provided the raw input data in CSV format; Matlab outputs are exported to the NetCDF format for reading into Scala.

We use as a background a map [11] of the 48 contiguous United States obtained from Wikimedia Commons. Points on the map correspond to real-world $(latitude, longitude)$ coordinates according to a mathematical transformation specified with the map; this allows us to draw physical locations at the correct positions on the map.

3.2 Graphical representation details

The visualization page consists of three components: an animation showing the dynamic status of the geographical load balancing system, a line plot showing the aggregate statistics of energy supply and demand, and a progress bar allowing progress scroll and control.

The animation displays 10 data centers on a base map of the United States. Each data center's status is represented by a sector diagram, which is animated over time. The left half of the circle represents the energy demand of the data center: the yellow sector is the energy demand for processing the request, the blue one the energy demand for cooling. The right half of the circle represents the energy supply: the light green sector represents available wind energy, the dark green one represents available solar energy, and the brown one represents the energy usage from the grid. The area of each sector is proportional to the amount of energy (so its radius is proportional to the square root). In addition, each sector diagram is surrounded by a faint circle, which represents the maximum energy usage when the data center operates at full load.

The request traffic $\lambda_{d,s}(t)$ is represented by lines connecting each source s and destination data center d. The width of the line is linearly proportional to $L_s(t)$, the total volume of requests from s. The transparency of the line is linearly proportional to $\lambda_{d,s}(t)/L_s(t)$, the percentage of the traffic from the source s routed to each data center. A solid black line means all the requests from s are sent to d, while a fully transparent line means that no traffic is routed from s to d.

The line plot displays four aggregate statistics of interest over time: the yellow line shows the aggregate IT power, the blue one the aggregate energy usage on cooling, the light green one the aggregate wind energy available, and the dark green one the aggregate solar energy available.

The progress bar at the bottom indicates the time elapsed in the animation. The user can pause/resume the animation and seek to the desired time.

4. RESULTS

4.1 Cost savings from Cooling-aware GLB

The cooling-aware geographical load balancing model reduces firm's energy cost by routing the requests to locations where energy is cheap and cooling is easy. Such a savings outweighs the increase in delay cost. Hence our model is consistent with profit maximization by a firm.

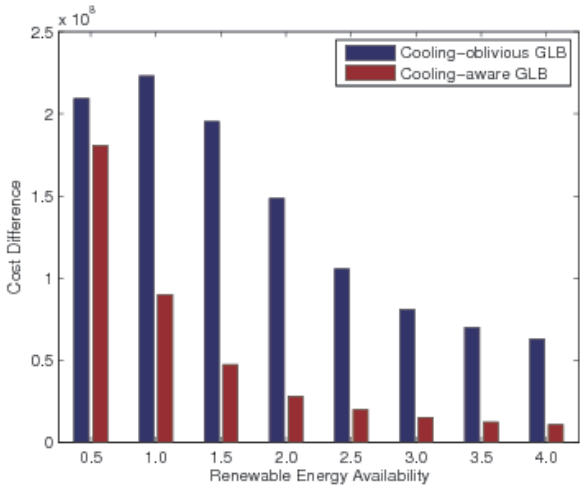


Figure 3: The difference between optimal costs in summer/winter for both Cooling-aware and Cooling-oblivious GLB

Our first experiment explores the cost efficiency of this model. Figure 2 illustrates the extent of total cost savings of our model over previous models, in summer and winter. We use as benchmarks the total costs under Cooling-oblivious GLB (the previous GLB optimization that does not consider cooling) and LOCAL (i.e. traditional routing without GLB, routing all requests to the nearest DC, as described in [13]). The Cooling-aware model gives lower cost than either benchmark. Its advantage is clearest in two situations.

First, in summer when cooling is expensive (Figure 2, left), the Cooling-aware model significantly outperforms the Cooling-oblivious model. By considering the energy cost of cooling, it makes routing decisions that better use renewable energy.

Second, when the aggregate renewable energy available lies between one and two times the aggregate demand, and some but not all data center locations have a surplus in renewables, the Cooling-aware GLB model exploits this surplus better than the benchmark models.

We are interested in how well the Cooling-aware GLB model performs under different seasons and weather conditions. Figure 3 shows the difference between the optimized costs in summer and in winter using each model. The previous Cooling-oblivious model's performance varies significantly with season; our Cooling-aware GLB model's performance is more robust. This finding reflects the fact that our Cooling-aware model exploits the heterogeneity in weather to reduce the energy required for cooling.

4.2 Environmental impact: CO_2 savings

The geographical load balancing model exploits variation in energy costs by routing requests to where energy cost is low. When the data centers have on-site renewable energy plants

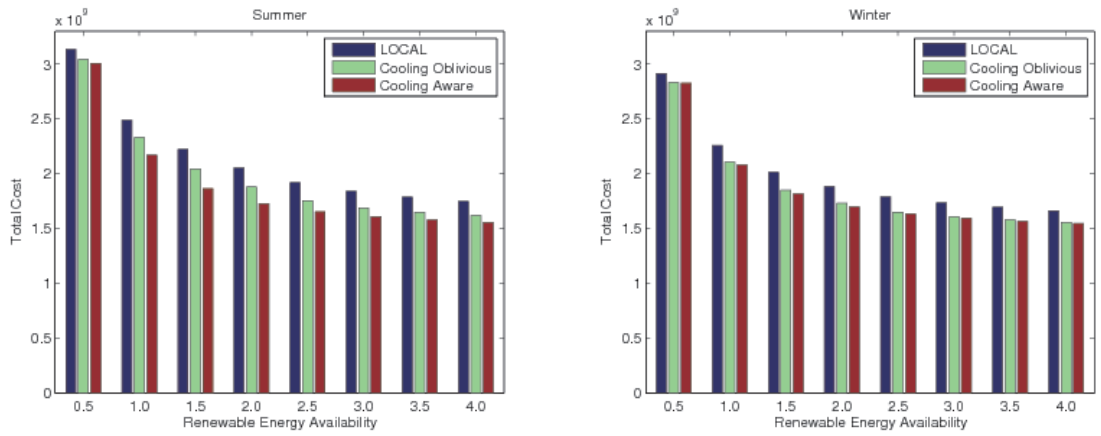


Figure 2: Comparison of optimal costs under Cooling-aware GLB, Cooling-oblivious GLB, and LOCAL, with varying renewable energy availability, in summer (left) and winter (right)

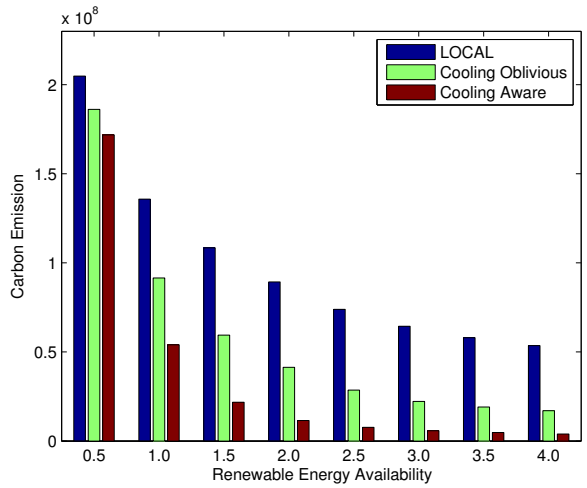


Figure 4: Comparison of carbon emissions under Cooling-aware GLB, Cooling-oblivious GLB, and LOCAL

(as in our setting), renewable energy becomes cheap; the model allows optimal usage of these renewables and thus has the desirable side-effect of reducing carbon emissions.

We can estimate the CO_2 emissions by multiplying the grid energy usage of each data center location by the CO_2 emissions per energy unit. Figure 4 shows the CO_2 emission comparison of each model in summer. By using the Cooling-aware GLB model, aggregate carbon emissions can be reduced by more than 50% as compared to LOCAL if the total renewable energy supply is equal to the total IT demand. If more renewable energy is available, carbon emissions under Cooling-aware GLB is less than half those of Cooling-oblivious GLB.

The environmental impact of our new model is significant. CO_2 emissions are reduced even when firms' optimizations do not consider the environmental externality of CO_2 emissions. Environmental benefits could be even greater if firms are given incentives to be environmentally friendly.

4.3 GLB versus storage

One alternative way of using renewable energy efficiently is to store renewables available in excess of current demand for future use. Whereas applying geographical load balancing adjusts energy demand according to supply in each location, adding energy storage adjusts supply according to demand. We are interested in the performance characteristics of both methods.

In this experiment, we compare the cost curve and the brown-energy usage curve of using GLB and using storage. Since our experiment starts at midnight, at which time a data center has likely used all of its storage for batch jobs [15], we assume that the storage starts empty.

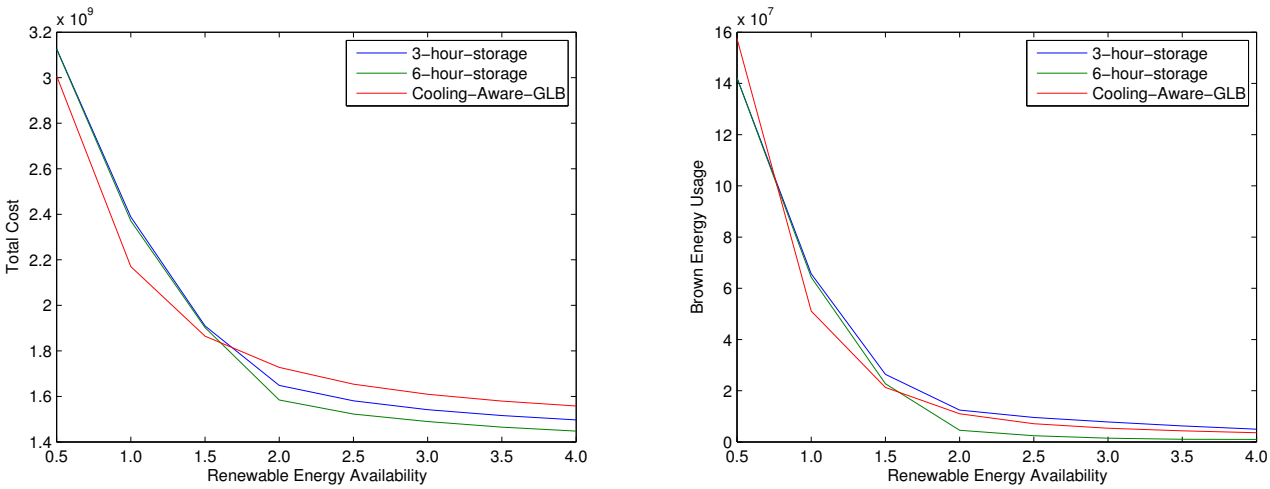
Figure 5a illustrates the trend of optimal costs under the two models with respect to varying renewable availability. It shows that Cooling-aware GLB has lower costs when the total renewable energy is less than 1.5 times the aggregate IT demand, but is outperformed by the storage model when renewable energy is in large surplus.

Figure 5b compares brown energy usage under each model. The result suggests that the Cooling-aware GLB model needs less energy from the grid compared to the 6-hour-storage curve when the total renewable supply is less than 1.5 times of the total IT power demand. Moreover, it needs less grid energy than the total renewable energy compared to the 3-hour-storage model, in almost the entire renewable availability interval, except when the renewable availability coefficient is 0.5. The environmental impact advantage of our new model is even more salient than the cost advantage.

We thus can conclude that our model is better than the storage model when the renewable generation facilities are not yet built, because it requires less prior infrastructure investment.

5. FUTURE WORK

Our simulation and visualization software can be used to answer questions about extensions to the model that would be useful when putting it into practice, such as—



(a) Cost

(b) Brown energy usage

Figure 5: Comparison of optimal cost (5a) and brown energy usage (5b) under the storage model and the Cooling-aware GLB model

- incorporating other kinds of renewables into the model
- treating data centers as endogenous. In the long run, where should we build new data centers, and what local renewables and energy storage systems do we use?
- finding the optimal control timescale. It is costly to switch servers on and off, so the current algorithms only change routing every fixed period, which is suboptimal. This significantly restricts the savings that follow-the-renewables routing can yield, but the topic hasn't yet been explored deeply. Being able to simulate the whole system helps for this.
- helping discover and then exploit emergent phenomena when looking at systems of many data centers. An example of such a phenomenon deals with the optimal mix of renewable energy for powering data centers. Since the sun shines in the daytime when people are awake and searching the internet, one might expect solar power to dominate. But previous work showed that in fact wind is more effective, because its high variability and low spatial and temporal correlation with demand is especially useful for geographical load balancing. With solar power, when requests come at night DCs are forced to buy electricity off the grid, but with wind power, it's almost always windy somewhere.

References

[1] In: (2010). URL: http://www.nrel.gov/rredc/.

[2] In: (2010). URL: http://wind.nrel.gov/.

[3] In: (2011). URL: http://www.datacenterknowledge.com/archives/2011/08/01/report-google-uses-about-900000-servers/.

[4] In: (2011). URL: http://www.dcd-intelligence.com/Industry-Census/Energy-Demand-2011-12.

[5] In: (2011). URL: http://thepowerfactor.wordpress.com/2011/04/21/co2-emissions-per-electrical-energy-unit-kgco2kwh-by-us-state/.

[6] In: (2012). URL: http://www.datacenterknowledge.com/archives/2012/05/15/google-data-center-faq/.

[7] In: (2012). URL: http://www.techweekeurope.co.uk/news/data-centre-power-increase-cloud-95359.

[8] In: (2012). URL: http://www.ncdc.noaa.gov/.

[9] In: (2013). URL: https://en.wikipedia.org/w/index.php?title=Data_center&oldid=540161748#Greenhouse_gas_emissions.

[10] Luiz André Barroso and Urs Hölzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan and Claypool, 2009.

[11] "File:Usa edcp location map.svg - Wikimedia Commons". In: (). URL: https://commons.wikimedia.org/wiki/File:Usa_edcp_location_map.svg (visited on 07/10/2012).

[12] Jonathan Koomey. "Growth in Data center electricity use 2005 to 2010". In: *Analytics Press* (2011). URL: http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion.pdf.

[13] Zhenhua Liu et al. "Geographical Load Balancing with Renewables". In: *Proceedings of ACM Greenmetrics*. 2011. URL: http://rsrg.cms.caltech.edu/greenIT/papers/geoloadrenewables.pdf.

[14] Zhenhua Liu et al. "Green Geographical Load Balancing". In: *Preprint* (2011).

[15] Zhenhua Liu et al. "Renewable and Cooling Aware Workload Management for Sustainable Data Centers". In: *Proceedings of ACM Sigmetrics*. Sigmetrics held jointly with IFIP Performance. 2012.