Global Marginal Carbon Footprint Evaluation of Internet Services with Building Energy Models

Authors List (Hidden for Review)
Institution (Hidden for Review)
Department (Hidden for Review)

Abstract

Globally distributed internet services are ubiquitous today. The network dependencies between physically dispersed data center resources make system level building design decisions nearly impossible to intuitively reason about. To support system level design decisions, this research demonstrates the operational carbon footprint analysis of globally distributed internet services using the service's network traffic profiles, data center building energy simulations, and utility grid-level marginal costs of energy models. The research presents a method to evaluate the environmental footprint attributed to data center buildings of global internet services that systems designers and operators can use to assess data center infrastructure.

Introduction

In a 2015 study data centers were forecast to consume as much as 13% of the global energy production by 2030 (Andrae and Edler, 2015). More optimistic models form the US Department of Energy for the US showed up to fifty percent decline in energy demand when using more innovative best practices relative to the industrys use of 2% in 2005 (Shehabi, 2016). As more and more parts of society are transitioning to data center dependent online paradigms, the absolute demand of data centers is growing. The volume of growth is exemplified by the capital costs being invested across the world in constructing these data centers.

The growth in data center capital construction costs will reach $89 billion by 2027, a significant increase from the $45 billion spent in 2018 (Insight-Partners, 2019). Capital costs arent the only commitments for data center owners however. DC owner's also incur significant operational costs throughout the entire operational lives of their facilities. Over a 20-year life of an continuously operating data-center facility, the use-phase energy costs can exceed its capital costs while having a much larger ecological footprint. Given the accumulation of costs and impacts over the life of data centers, there is a need for a robust model the couples the building systems with their energy supply sources. In this article a geographically extensible model that meets these needs is presented.

In the rest of this paper a model for coincident energy demand and marginal energy costs of a set of data centers is proposed and developed. The proposed method uses a hybrid model consisting of Energy Plus and Python programming modules as developed by the researcher (Kumar, 2020a). The original BEM is modified as described in the methodology section to be more indicative of modern cloud data centers in this work. The resulting time series energy demand profile is then passed as input to a marginal energy cost simulation tool that is also extended as part of this work.

In this work the marginal costs of energy are calculated using the Dispatch Optimized Systems Cost of Energy model developed by Platt (Platt et al., 2017). DOSCOE provides a linear programming platform that computes the monetary costs and several environmental emissions associated with operating a mix of dispatch-able and non-dispatch-able energy sources. In this context, dispatch-able energy sources are those that can be controlled to meet demand. Natural gas power plants are examples of dispatch-able source of energy, where the plant operators can control the mass flow rates of the combustion gases to curtail generation rates. On the contrary are the non-dispatchable sources, where the energy generation is dependent on extrinsic factors. An example of non-dispatchable energy source is solar; where cloud cover greatly effects the generation rate and no power can be generated at night.

Modeling a mixture of dispatch-able and non dispatchable generators is complex as most non-dispatchable sources are only accounted for as opportunistic supply sources when sizing the power generation infrastructure. While the dispatch-able generators are sized to meet full demand in a worst-case condition when no dispatch-able power is available (Platt et al., 2017). When non-dispatchable power is

opportunistically available, it supplements the grid, allowing dispatch-able sources to be turned down to part loads. This saves the fuel costs for dispatch's sources (ie natural gas generators), but it leaves the physical infrastructure stranded and underutilized. This researchs coupling of DOSCOE model with a BEM allows the monetary and ecological costs associated with operating data centers in grids with mixed energy sources to be evaluated by matching the marginal costs of energy with data center demands on hourly intervals.

In the next section background context is provided to set the proper use case for this framework. Then in the similar past works section, past literature that have quantified the ecological life cycle cost of data centers and internet services are summarized. In the methodology section, the two main modules of the software implementation of this research are presented. Specifically these modules are a modified version of hybrid building energy model and a novel marginal energy cost model based on DOSCOE. After the details of the model are presented, the results are discussed in the discussion section followed by the conclusion.

Background

Data centers are critical part of the modern internet experiences for billions of people. They are the key enablers for disseminating information in real time regardless of people's location. Figure 1 shows an example of a logical architecture that enables data centers to provide globally consistent internet experience that have become the status quo over the last decade. In Figure 1, two layers of hierarchy in the global data center stack are shown.

The first hierarchical layer, the global level, has a wide area network (WAN) connected to two internet service providers (ISP) as its root node. For the purposes of this work the WAN is an abstraction of a network that connects a set of data center facilities with each other. The second layer of the global level are the metropolitan regions. In this layer, the ISP links are shown as coming in from top and the outgoing links from the bottom serve local distribution networks of metropolitan area facilities. The final layer of the global level are the data center facilities, where the global network links connect to clusters of servers. Inside each data center there may be another independent hierarchy.

Large cloud data center operators have championed WAN systems for global up time and user experience optimizations (Sushant et al., 2013) (Hong and Kandula, 2013). Cloud data centers are a specific class of data centers that are comprised of custom information technology and building systems tuned for optimal total costs of ownership based on proprietary models at best or they optimized simple based on

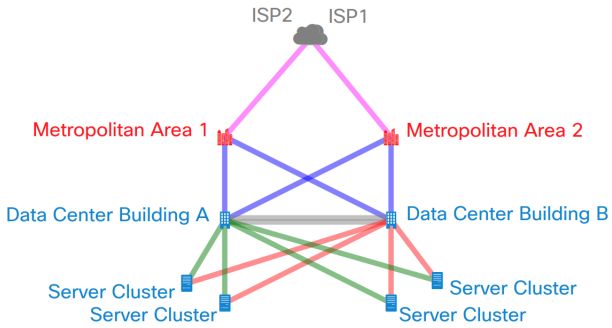

Figure 1: General Topology of data center networks. The one to one metropolitan area to building relationships are shown for clarity only.

heuristics. These facilities house numerous services that can be controlled by network load balancers, with each data center limited by its physical infrastructure's capacity. Given the applications described by Sushant and Hong, the network technology can shift loads on command, but doing so incurs additional costs for having hot spares. An effective hot spares strategy requires fully redundant replicas that, by design, would sit idle when other facilities are in operations.

As noted by the enormous market value of data centers, businesses need robust models that can be used to assess the trade-offs between network flexibility and the physical infrastructure. The fungibility of locations enabled by the network technologies is also desirable from an application performance point of view as latency can be significantly reduced for specific markets and applications by reducing the round trip communication times with consumers. Given the popularity of the public clouds, the application of network aware energy model for data centers extends beyond just data center construction and plant operations.

In the context of building energy modeling, WAN's make the information technology workloads temporally (and geographically) unstable, rendering it elusively for building design professionals to reason about. As an example of temporal and geographical instability of data centers, Figure 2 (Barroso et al., 2018) shows the difference in utilization rates for two server clusters from their operational experience. The x-axis indicates the cumulative time ratio and the y-axis shows the utilization. The figure on the left shows that the peak utilization only happens for 40% of the operational time. This is an area that building level energy simulations can help exploit. However, building energy simulation need to be aware of more then just energy. They must also be capable of quantifying the energy's global warming potential in terms of carbon footprint.

Based on the literature reviews and references cited,

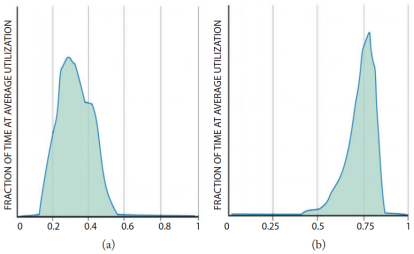

Figure 2: Average activity distribution of a sample of 2 clusters, each containing over 20,000 servers, over a period of 3 months. (Barroso et al., 2018)

Figure 3: Data center locations

there are no publicly available simulation frameworks that couple the dynamic data center workloads and the coincident carbon footprint associated with powering their load. The marginal cost of energy coupled building energy model described in this research is the first publicly available tool to allow owners, designers, and researchers to quantify the carbon footprint of data center operations accounting for granular supply and demand matching of power. Next, a literature review of past works concerning the carbon footprint of data-centers and digital services is presented in the Similar Works section.

Similar Works

There are two notable past works that look at the energy footprint for distributed sets of data centers. First, Tripadi considers hardware capital costs alongside with energy acquisition costs to quantify the total costs of ownership in (Tripadi et al., 2017). Tripadi's framework is dynamic in terms of workloads but it is not aware of the building energy dynamics. In the second work, Kiani and Ansari describe a geographical load balancing strategy that exploits green energy mix in the utility grid (Kiani and Ansari, 2017). However, their load balancing scheme doesn't provide insights into how the load balancer evaluates the building energy demands or how the greenness of the energy supply is obtained.

Using a life cycle assessment framework, Whitehead demonstrates a comprehensive data center site level life cycle costs analysis in (Whitehead, 2015). All energy use in Whiteheads models were deduced from annualized PUE values, precluding coincident energy source evaluations with their framework. Similarly the Green Grids data center life cycle assessment guideline is limited to PUE as their suggested basis for the operational energy proxy (The Green Grid, 2012).

Others have evaluated the costs of internet services (Taylor and Koomey, 2008) (Shehabi et al., 2014). Taylor and Koomey quantify the energy and greenhouse gas implications of online Advertising circa 2008. While Shehabi evaluates the energy and greenhouse gas implications of video stream circa 2014 (Shehabi et al., 2014). Together these works provide a taxonomy that can be followed to assess internet service costs in terms of energy use.

Methodology

In this section the marginal energy cost model's coupling with the BEM is described. The resulting model maintains a strict partition between the two technical domains. The first part of the model simulates the hourly energy demands of a set of five data centers in EnergyPlus (EP) using the method demonstrated in (Kumar, 2020a). These data centers are distributed across the globe as shown in Figure 3. Then in the second part, the data center demands are matched with the respective region's utility power generation sources for the coincident hour to assess the marginal energy cost during that hour. These parts are described in the following subsections.

Building Energy Model

For the first part, a sufficient building energy demand profile is simulated as in (Kumar, 2020a). The profile is obtained by using 50th quantile of the daily network traffic to each data center using Kumars method from (Kumar, 2020b). However in lieu of using EP release 8.6, the simulation in this work uses EP release 8.9. The changes found EP 8.9 are listed in NRELs GitHub (NREL, 2020). The programmatic changes from EPv8.9 resulted in three notable differences between original BEM and the model presented here. The first change was motivated by original BEM's runtime errors when setting $2kW/m^2$ as the information technology (IT) equipment load density. This persistently led to thermal runaway conditions for the Singapore and San Francisco sites in EP release 8.9. In order to keep the building envelope form-factor and the construction materials the same, this simulations IT power load density is set to $1.0kW/m^2$.

The second change was made to exploit a new feature for modeling supply and return air compartmentalization in EP release 8.9. In these simulations, the data center model has air distribution flow control with approach temperatures specified. Flow control with approach temperature method calculates the temperature differences between the IT hardware boundaries

and the air handling equipment. This method is more representative of modern data-center operations and allows modeling ASHRAE 90.4 Standards requirement of hot-aisle / cold-aisle compartmentalization. The alternate method in EnergyPlus considers the entire data center room as a well-mixed environment, consistent with the modeling from (Kumar, 2020a). While using the approach temperature method, the cooling temperature is changed for each time-step to 27 degrees Celsius from 25 degrees used in (Kumar, 2020a). This 2-degree adjustment corresponds with the approach temperature between the air handling equipment discharge and the inlets of the ITE.

As the third change, the load distribution specification in the IDF was revised from the SequentialLoad setting to UniformPLR. In the former, equipment is activated in the order listed in the IDF. With this specification each piece of equipment ramps sequentially from its idle state to full capacity, before subsequent equipment is enabled. In the latter load distribution specification, all equipment are loaded in parallel to each other. Another available setting for load distribution, the Optimal specification, was also tried for this field, but it crashed the simulations.

To validate that the proposed building energy model configurations are sufficient, with the above changes, each data-center is simulated with two models. In the first model, the economizer for the direct evaporative cooler (DEC) limits are relaxed while in the second model the default settings are maintained. The summary of the changes are indicated in Table 1, while the comparison of total site energy between the two models for each set of the simulations is illustrated in Figure 4. A site's pair simulations corresponds to the bars in the figure where the same data center is modeled with varying IDFs. In the figure the green bar indicates the total site energy of the model with economizer dry-bulb temperature limits relaxed. From the figure it is observed that the relaxed economizer leads to more energy use than the original values. While it maybe possible to fine optimal economizer settings, for this article it's concluded that the defaults settings from Kumar's original work suffice.

Table 1: Relaxed economizer settings

IDF Object	Variable	Original/Econ
DC-OA	Econ. Max db-C	23 / 28
DC DEC	Evap. Max db-C	20 / 28

The next subsection introduces the marginal costs of energy model,which will take the results from the original BEM.

Regional Marginal Costs of Energy

The second module requires a time-series profile of the electrical grids power source attributions, where the time steps between the data center energy use profile match with the energy generation. The corresponding site and regional grid model pairs are indicated in Table 2.For this model the energy generation regional grid profiles are obtained from Platts DOSCOE [5]. Singapore and Amsterdam datacenters grid profiles are constructed as described below.

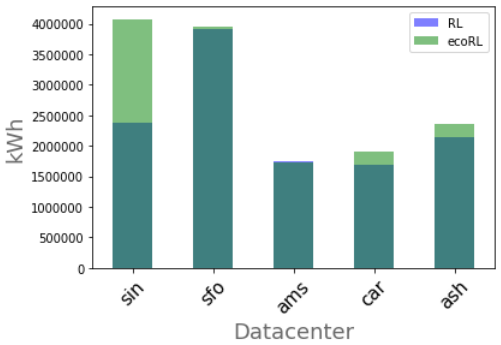

Figure 4: Comparison between relaxed economizer settings vs. default IDF settings from (Kumar, 2020a)

Table 2: Data Center Site and Power Grid

DC ID	DC Location	Regional Grid
SFO	San Francisco, CA	California
CAR	Carrolton, Texas	ERCOT
ASH	Ashburn, VA	Midatlantic
AMS	Amsterdam	Netherlands
SIN	Singapore	Singapore

Each U.S region represented in the grid profiles is composed of hourly demands and corresponding production capacity of several power generation technologies. Hourly values for each regions demand, solar, wind, coal, coal, coal with cryogenic capture, coal amine gas scrubbing, nuclear are defined. These grid regions are representative of three out of five of the data-center locations being simulated.

For Singapore, the International Energy Agency (IEA) provides a top down view of the annual energy production from various technologies. The IEA data shows that the renewable penetration in the energy supply for Singapore accounted for only 1.6% of the total energy demand in 2016,the last year the data is available for (IEA, 2017). Due to this negligible contribution from renewable sources and lack of hourly generation data, it is assumed that all demand is met by natural gas power generators, consistent with other sources (EIA, 2016b).The demand profile for Singapore in 2016 is obtained from (Singapore-Governement, 2016).

For the Amsterdam data center, energy profile of Netherlands is used. The IEA indicates that in 2018, 12% of Netherlands's energy demands was met by renewable sources. This is a meaningful contribution from renewable, therefore a more granular generation profile is prudent. To formulate the granular resolution of the generation, the IEA data is supplemented with data from the Open-Power-System-Data (Open-Power-System-Data, 2019). However, OPSD provides

hourly data for the renewable sources only. The balance of the energy demands in the Netherlands is met by fossil fuel-based generators, namely 35% by coal and 42% by natural gas (EIA, 2016a). The DOSCOE grid profiles dont have any corresponding field for bio-mass, so the bio-mass generation indicated in OPSD is lumped in with the fossil fuel generators. In the discussion section, some validation for this approach is presented. The Netherlands also produces nuclear energy, but only the annual production rates were obtained (EIA, 2016a). The annual nuclear production is distributed equally over the year and modeled as a constant (non-dispatchable) base load throughout the year.

The MEC coupled BEM algorithm is given below. In the algorithm two inputs are required. The first is DOSCOE[region], it is a two-dimensional array formatted as described in (Platt et al., 2017). It indicates hourly grid profiles of the power demand and power capacity of the available energy sources at the corresponding hour. The second inputs, traffic.language, is a one-dimensional vector indicating the network traffic to the respective site. Using the second input, for each language, (the proxy for an internet service) its traffic to the respective data center is checked. If there are traces of a language routed to a site, the algorithm performs the BEM simulation by invoking EnergyPlus. This resulting demands from EnergyPlus are then added to the regions grid demand profile. If a language does not have any traffic to a particular site than, the data center site does not do any work and the BEM simulation is bypassed.

Algorithm 1 MEC coupled BEM algorithm

Require: DOSCOE[region] & traffic.language[site]
  for site and region in DC.site and DC.region **do**
    if traffic.language[site].all ! = 0: **then**
      $D_{DC} \leftarrow BEM(site, traffic.language)$
      $DOSCOE[region].demand \mathrel{+}= demand[site]$
    end if
  end for
  CO_2 footprint = $GridSim(region, rps)$
  Where D = $DEMAND_{DC}$

The final step of algorithm quantifies the carbon footprint of the grid with the added data center loads by running DOSCOEs Grid Simulator. In this step, the renewable portfolio standard (rps) argument specifies the percentage of renewable energy mix for the region.

In the next section the results from the methodology are discussed.

Discussion

The resulting values for source energy carbon footprint for each data center is summarized in Table 3 below. The energy model for the 491kW data center has been scaled by 1000 to represent the metro scale of data centers. The table further indicates the carbon footprint of each of the languages being served from the data centers. The values are indicative of the marginal amount of CO2 emitted by adding the data center demands to the grid. The units of the values are in Tons of CO_2 emitted.

From Table 3 indicates the total marginal carbon footprint for each language by summing the language column. While the total marginal carbon footprint of each data center is obtained by summing the rows. The English pages have the most traffic globally and as expected English has the largest marginal carbon footprint due to its data-center operations. It is surprising however that the Netherlands data center has the highest total marginal carbon footprint among the data centers. A relatively lower marginal costs is expected due to the comparatively cooler ambient temperature in the Netherlands which would allow the data center systems to use economizers for more hours, but given that many languages are routed to the Amsterdam data center makes its absolute workload relatively is higher than other data centers. The EnergyPlus results support the lower building energy demands for the Netherlands site as shown in the stacked PUE histogram in 5. With the PUE values indicating the efficiency of the of the Netherlands site being more than average, the marginal carbon footprint can be attributed to the energy supply mix and higher network traffic from multiple languages. The aggregated the grid profile as described above for DOSCOE is heavily biased towards fossil fuels, therefore its carbon emissions are higher. This work used several secondary sources to construct the grid profile for the Netherlands and these sources may need to be validated more rigorously in the future.

As an implementation detail, the DOSCOE model proved to be quite sensitive to the data structure of the load profiles. For example any null value in the profile resulted in breaking the linear solver. Also, Solar and wind energy are required to be non-zero for at least one hour of the year. This requirement is a practical constraint, but in the profile developed for this work the values are arbitrarily set to a low value across all hours. Specifically, the defaults setting used in the work is 0.1 MW for each hour.

Conclusion

The network dependencies between physically dispersed data-center resources make system level building design decision a challenge to reason about. This research has presented a quantitative method to couple the network dependencies of data centers with their building energy and grid level marginal costs of energy. As shown for the Netherlands site, lower building level efficiencies, such has PUE does not equate to lower carbon footprint.

As pointed out in the background section, there is a lack of bottoms up building design and carbon foot-

Table 3: Data Center Carbon Footprint

DC ID	de	en	es	fr	ja	ru	zh	total
SFO	0	1,301,549	0	0	0	0	0	**1,301,549**
CAR	0	519,466	1,341,227	0	0	0	0	**1,860,693**
ASH	0	563,085	1,109,162	661,919	0	0	0	**2,334,166**
AMS	464,780	470,714	473,987	766,542	0	1,027,424	457,393	**3,660,840**
SIN	464,780	470,714	473,987	766,542	0	1,027,424	457,393	**3,660,840**
Sum	**464,780**	**3,446,114**	**2,924,376**	**2,042,611**	**873,602**	**1,888,019**	**1,035,892**	**12,675,394**

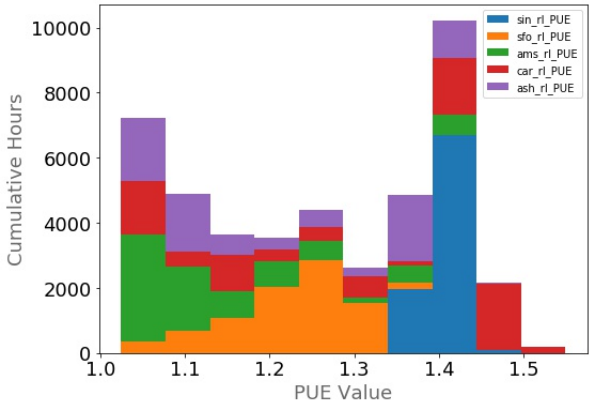

Figure 5: Stacked histogram of PUE from all Data Centers.

print models. The modular methodology of this research is a novel means of coupling abstracted service level metric, WAN traffic in this case, with physical based building energy simulation, EnergyPlus in this case. This integrated tool set can be used by data-center designers and operators to optimize their deployment across geographical bounds. Future work should consider controlling the network traffic based on building energy simulations to achieve global optimality in real time operational environments.

References

Andrae, A. and T. Edler (2015). On global electricity usage of communication technolgy: Trends to 2030. In *Challenges*, Volume 6, pp. 117–157.

Barroso, L. A., U. Hozle, and P. Ranganathan (2018). *The Datacenter as a Computer: Designing Warehouse-Scale Machines 3rd ed.* Morgan and Claypool.

EIA, U. (2016a). U.S. Energy Information Administration. Retrieved: May 02, 2020. wwww.eia.gov/international/analysis/country/NLD.

EIA, U. (2016b). U.S. Energy Information Administration. Retrieved: May 05, 2020. wwww.eia.gov/international/analysis/country/SGP.

Hong, C.-Y. and S. Kandula (2013). Achieving high utilization with software-driven wan. In *SIGCOMM'13 ACM, Hong Kong.*

IEA (2017). International energy data and statistics. Retrieved: March 03, 2020.

Insight-Partners (2019). Data center construction market. Retrieved: May 12, 2020. https://www.theinsightpartners.com/reports/data-center-construction-market.

Kiani, A. and N. Ansari (2017). On the fundamental energy trade-offs of geographical load balancing. Volume 0163-6804/17, pp. 170–175.

Kumar, E. (2020a). Towards energy simulations for proportionally designed and controlled data centers. In *Accepted for Building Simulation and Optimization 2020.* Loughborough (UK), 21-23 September 2020.

Kumar, E. (2020b). Wide area network based data-center energy simulations for internet services. In *Accepted for 11th International Conference on Improving Energy Efficiency in Commercial Buildings and Smart Communities.* Frankfurt (Germany), October, 2020.

NREL, N. R. E. L. (2020). Nrel GIthub energy plus. Retrieved: January 02, 2020. //github.com/NREL/EnergyPlus/releases.

Open-Power-System-Data (2019). Time series open-power-system-data. Retrieved: April 02, 2020. https://data.openpowersystem-data.org/time_series/20190605.

Platt, J., O. Pritchard, and D. Bryant (2017). Data center construction market 2019-2024 — global industry overview by size, share, future demand, latest research by competitors, segmentation and regional forecast. *SSRN 08*.

Shehabi, A. (2016). United States Data Center Energy Usage, United States Department of Energy.

Shehabi, A., B. Walker, and E. Masanet (2014, may). The energy and greenhouse-gas implications of internet video streaming in the united states. *Environmental Research Letters 9*(5).

(2016). *Half Hourly System Demand. Retrieved: January 02, 2020*. https://data.gov.sg/dataset.

Sushant, J., S. Mandal, J. Ong, A. Singh, U. Holzle, and A. Vahdat (2013). B4: Experience with a globally-deployed software defined WAN. In *SIGCOMM'13 ACM, Hong Kong*.

Taylor, C. and J. Koomey (2008). Estimating the use and greenhouse emissions of internet advertising, IMC squared.

The Green Grid, T. (2012). Data center life cycle assessment guidelines. Retrieved: May 13, 2020. https://www.thegreengrid.org/en/resources/library-and-tools/236-Data-Center-Life-Cycle-Assessment-Guidelines.

Tripadi, R., S. Vignesh, V. Tamarapalli, and D. Medhi (2017). Cost efficient design of fault tolerant geo distributed datacenters. In *IEEE Transactions on Network and Service Management*.

Whitehead, B. (2015). The life cycle assessment of a UK data centre. In *International Journal of Life Cycle Assessments*, Volume 20, pp. 332–349.