Chapter 2
Energy and Carbon Footprint-Aware Management of Geo-Distributed Cloud Data Centers: A Taxonomy, State of the Art, and Future Directions

Atefeh Khosravi
The University of Melbourne, Australia

Rajkumar Buyya
The University of Melbourne, Australia

ABSTRACT

Cloud computing provides on-demand access to computing resources for users across the world. It offers services on a pay-as-you-go model through data center sites that are scattered across diverse geographies. However, cloud data centers consume huge amount of electricity and leave high amount of carbon footprint in the ecosystem. This makes data centers responsible for 2% of the global CO2 emission. Therefore, having energy and carbon-efficient techniques for resource management in distributed cloud data centers is inevitable. This chapter presents a taxonomy and classifies the existing research works based on their target system, objective, and the technique they use for resource management in achieving a green cloud computing environment. Finally, it discusses how each work addresses the issue of energy and carbon-efficiency and also provides an insight into future directions.

INTRODUCTION

In recent years the use of services that utilize cloud computing systems has increased greatly. The technology used in cloud is not new and its main goal is to deliver computing as a utility to users. Cloud computing consists of virtualized computing resources inter-connected through a network, including private networks and the Internet. Over the years since its formation, different definitions for cloud

DOI: 10.4018/978-1-5225-2013-9.ch002

computing have been proposed. According to the definition by the National Institute of Standards and Technology (NIST) (Mell and Grance, 2011): "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models". The three service models provided by the cloud providers are Infrastructure, Platform, and Software as a Service.

Cloud computing delivers service, platform, and infrastructure services to users through virtual machines deployed on the physical servers. Virtualization technology maximizes the use of hardware infrastructure and physical resources. Hardware resources are the servers located within the data centers. Data centers are distributed across the world to provide on-demand access for different businesses. Due to the distributed nature of cloud data centers, many enterprises are able to deploy their applications, such as different services, storage, and database, in cloud environments. By the increase of demand for different services, the number of data centers increases as well; which results in significant increase in energy consumption. According to Koomey (2008) energy usage by data centers increased by 16% from the year 2000 to year 2005. Energy consumption of data centers almost doubled during these five years, 0.5% and 1% of total world energy consumption in 2000 and 2005, respectively. Hence, during the recent years there has been a great work on reducing power and energy consumption of data centers and cloud computing systems. Recently, considering data centers carbon-efficiency and techniques that investigate cloud data centers energy sources, carbon footprint rate, and energy ratings have attracted lots of attention as well. The main reasons for considering carbon-efficient techniques are increase in global CO_2 and keeping the global temperature rise below 2°C before the year 2020 (Baer, 2008).

In the rest of the chapter, the authors provide an in-depth analysis of the works on energy and carbon-efficient resource management approaches in cloud data centers, based on the taxonomy showed in Figure 1. The authors explore each category and survey the works that have been done in these areas. A summary of all the works is given in Table 1.

Figure 1. Taxonomy of energy and carbon-efficient cloud computing data centers

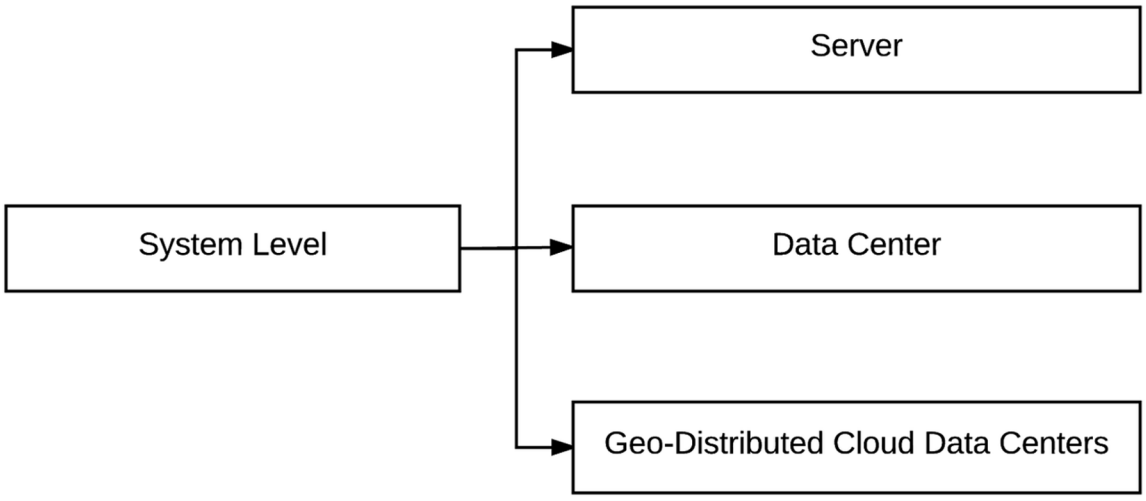


Table 1. Summary of various techniques for energy and carbon-efficient resource management in cloud data centers

Project Name	Goal	Architecture	Technique	Carbon-Aware
Dynamic right-sizing on-line algorithm, Lin. et al. (2011)	Minimize energy consumption and total cost	Single data center	Online prediction algorithms for the number of required servers for the incoming workload	No
Green open cloud framework, Lefevre et al. (2010)	Minimize energy consumption	Single data center	Predict the number of switched-on servers through providing in-advance reservation for users	No
Prediction-based Algorithms, Aksanli et al. (2011)	Maximize renewable energy usage and minimize number of job cancellation	Single data center	Use prediction-based algorithms to run the tasks (mainly batch jobs) in the presence of renewable energies	Yes
GreenSlot scheduler, Goiri et al. (2011)	Maximize renewable energy usage and minimize cost of using brown energies	Single data center	Prediction-based algorithms for the availability of solar energy and suspending the batch jobs in the absence of green energy	Yes
Multi-dimensional energy-efficient resource allocation (MERA) algorithm, Goudarzi et al. (2012)	Minimize energy consumption and maximize servers' utilization	Single data center	VM placement heuristic to split the VMs and place them on a server with the least energy consumption	No
Multi-objective VM placement, Xu et al. (2010)	Minimize power consumption, resource wastage, and the maximum temperature on the servers	Single data center	Data center global controller places the VMs based on a multi-objective algorithm to provide balance between power consumption and temperature	No
Green SLA service class, Haque et al. (2013)	Explicit SLA to guarantee a minimum renewable energy usage to run the workload	Single data center	Power distribution infrastructure to support the service and optimization based policies to maximize cloud provider's profit while meeting user's green SLA requirements	Yes
Cost-aware VM placement problem (CAVP), Chen et al. (2013)	Minimize the operating cost	Distributed data centers	VM Placement using meta-heuristic algorithms, considering different electricity prices and WAN communication cost	No
Energy model for request mapping, Qureshi et al. (2009)	Minimize electricity cost	Distributed data centers	Request routing to data centers with lower energy price using geographical and temporal variations	No
Free Lunch architecture, Akoush et al. (2011)	Maximize renewable energy consumption	Distributed data centers	VM migration and execution between data center sites considering renewable energy availability	Yes
Energy and carbon-efficient cloud architecture, Khosravi et al. (2013)	Minimize carbon footprint and energy consumption	Distributed data centers	VM placement heuristics to place the VM on the data center/cluster with the least carbon footprint and energy consumption and on the server with the least increase in power consumption	Yes
Framework for load distribution across data centers, Le et al. (2009)	Minimize brown energy consumption and cost	Distributed data centers	User request is submitted to the data center with access to the green energy source and least electricity price	Yes
Geographical load balancing (GLB) algorithm, Liu et al. (2011)	Minimize brown energy consumption	Distributed data centers	Use the optimal mix of renewable energies (solar/wind) and energy storage in data centers to eliminate brown energy consumption	Yes
Online global load balancing algorithms, Lin et al. (2012)	Minimize brown energy consumption and cost	Distributed data centers	Route requests to the data centers with available renewable energy using online algorithms	Yes
GreenWare middleware, Zhang et al. (2011)	Maximize renewable energy usage	Distributed data centers	Submit the requests to the data center site with available renewable energy, while meeting provider's budget cost constraint	Yes

continued on following page

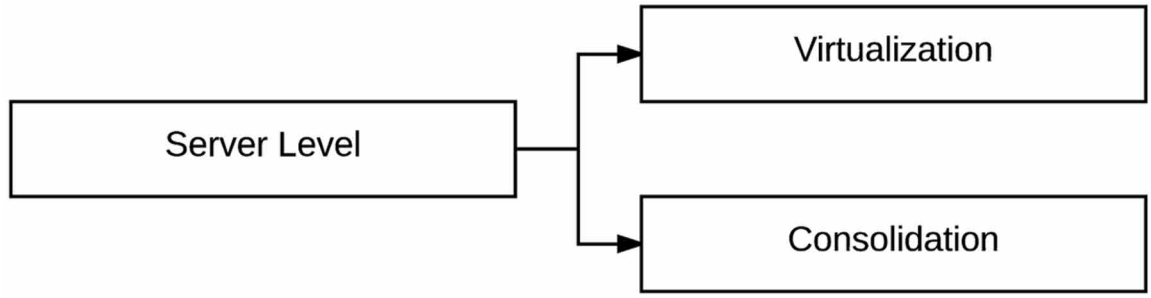
Table 1. Continued

Project Name	Goal	Architecture	Technique	Carbon-Aware
Environment-conscious meta-scheduler, Garg et al. (2011)	Minimize carbon emission and maximize cloud provider profit	Distributed data centers	Near-optimal scheduling policies to send HPC applications to the data center with the least carbon emission and maximum profit, considering application deadline	Yes
Carbon-aware green cloud architecture, Garg et al. (2011)	Minimize energy consumption and carbon footprint	Distributed data centers	Submit the user requests to the data center with the least carbon footprint, considering user deadline	Yes
MinBrown workload scheduling algorithm, Chen et al. (2012)	Minimize brown energy consumption	Distributed data centers	Copy the data in all the data centers, then based on the request deadline and the data center with least brown energy consumption executes the request	Yes
Federated CLEVER-based cloud environment, Celesti et al. (2013)	Minimize brown energy consumption and cost	Distributed data centers	Allocate the VM request to the cloud data center with the highest amount of photovoltaic energy and lowest cost	Yes
Temperature-aware workload management, Xu et al. (2013)	Minimize cooling energy and energy cost	Distributed data centers	Joint optimization of reducing cooling energy by routing requests to the site with lower ambient temperature and dynamic resource allocation of batch workloads due to their elastic nature	No
Provably-efficient on-line algorithm (GreFar), Ren et al. (2012)	Minimize energy cost	Distributed data centers	Use servers' energy efficiency information and places with low electricity prices to schedule batch jobs and if necessary suspending the jobs	No
Optimization-based framework, Le et al. (2010)	Minimize cost and brown energy consumption	Distributed data centers	Distribute the Internet services to the data centers considering different electricity prices, data center location with different time zones, and access to green energy sources	No
Dynamic load distribution policies and cooling strategies, Le et al. (2011)	Minimize cost	Distributed data centers	Intelligent placement of the VM requests to the data centers considering data centers geographical location, time zone, energy price, peak power charges, and cooling system energy consumption	No
Online job-migration, Buchbinder et al. (2011)	Minimize cost	Distributed data centers	On-line migration of running jobs to the data center with lowest energy price, while considering transport network costs	No
Spatio-temporal load balancing, Luo et al. (2015)	Minimize cost	Distributed data centers	Route the incoming requests to the data centers considering spatial and temporal variation of electricity price	No
Data centers' intelligent placement, Goiri et al. (2011)	Minimize cost, energy consumption, and carbon footprint	Distributed data centers	Find the best location for data center, considering location dependent and data center characteristics data	Yes
GreenNebula, a prototype for VM placement that follows-the-renewables, Berral et al. (2014)	Minimizing data center and renewable power plant building costs	Distributed data centers	Find the best geographical location to build data centers and renewable power plants and migrate the VMs, whenever necessary, to use a certain amount of renewables (solar or wind)	Yes

ENERGY EFFICIENCY IN SERVERS

Servers are the physical machines that run the services requested by users on a network. Servers are placed in a rack and any number of racks can be used to build a data center. Servers along with cooling systems and other electrical devices in the data centers consume 1.1-1.5% of the global electricity usage (Koomey, 2007). Hence, power and energy management of servers by the increase in users' demand for computing resources is irrefutable. Figure 2 shows a classification of techniques that are used in

Figure 2. Server level energy and carbon-efficient techniques



data center servers to reduce energy consumption. Virtualization and consolidation are two well-known strategies that make the data center servers energy-aware. These are two powerful tools that are applied in cloud data center servers in order to reduce energy consumption and accordingly carbon footprint.

Virtualization

Virtualization technology is the main feature of data center servers that leads to less energy consumption (Brey & Lamers, 2009). By having virtualized servers and resources, and using virtualization technology several virtual machines (VMs) can be built on one physical resource. Three types of virtualization that are widely used in data centers are hardware, software, and operating system virtual machines. The VMs run on the servers share the hardware components, that helps the operators to maximize server's utilization and benefit from the unused capacity. By maximizing server's utilization, huge savings in cost and energy consumption of data centers will be made. Decrease in data centers costs and energy consumption is not the only advantage of using virtualization technology. As the average life expectancy of a server is between three to five years, data and applications need to be consolidated and migrated to another server. Virtualization helps these two techniques to be done faster and with less cost and energy.

Consolidation

Server consolidation technique benefits from emerging of multi-core CPUs and virtualization technology. It's aim is to make efficient usage of computing resources to reduce data centers cost and energy consumption (Srikantaiah, Kansal, & Zhao, 2008). Consolidation is used when the utilization of servers is less than the cost associated to run the workloads (energy cost to run servers and cooling cost for data center servers). By using consolidation, servers can combine several number of running VMs and workloads from different servers and allocate them on a certain number of physical servers. Therefore, they can power-off or change the performance-level of the rest of physical servers and reduce the energy consumption, cost, and carbon footprint.

ENERGY EFFICIENCY IN DATA CENTERS

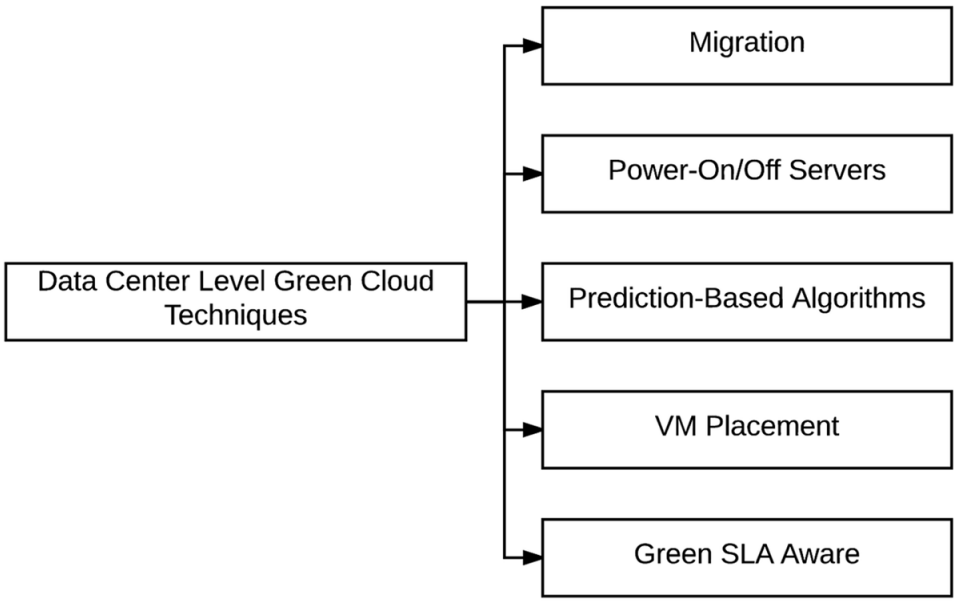
This section gives an overview of the researches that have been done at data center level to improve carbon and energy-efficiency of cloud data centers. An extensive taxonomy and survey of these techniques is done by Beloglazov, Buyya, Lee, and Zomaya (2011). Most of the works within a data center focuses on reducing energy consumption, which can indirectly result in carbon footprint reduction as well. Figure 3 classifies different approaches that have been taken for single data center. Some approaches use server level techniques (virtualization and consolidation) to migrate the current workload (user applications or virtual machines) and turn-off unused servers. Moreover, a provider could use the incoming workload pattern to place user request in the best suited cluster and server (and virtual machine for user applications) with less increase in energy consumption and carbon footprint.

Migration

Using virtualization, data center workloads migrate between servers. VM migration is the process of moving a running virtual machine from its current physical machine to another physical machine. Migration should be done in a way that all the changes be transparent to the user and the only change that user may encounter is a small increase in latency for the running VM or application.

Migration allows a virtual machine to be moved to another physical server so that the source physical server could be switched off or be moved to a power saving mode in order to reduce the energy consumption. VM migration in cloud data centers could be done off-line or live (Harney, Goasguen, Martin, Murphy, & Westall, 2007). There has been a great amount of work done in this area try to identify the VMs on the servers with low utilization that could be migrated, so that the provider can put the unused servers in idle or power-off state.

Figure 3. Data center level energy and carbon-efficient techniques



Power-On and Off Servers

When in an idle state data centers consume around half the power of their peak utilization and power state (Barroso & Holzle, 2007). There are technologies that try to design data center servers so that they just consume power in the presence of load, otherwise they go to a power saving mode. Work that is done by Lin, Wierman, Andrew, and Thereska (2011) uses a dynamic right-sizing on-line algorithm to predict the number of active servers that is needed by the arriving workload to the data center. Based on the experiments that are done in Lin et al. (2011) dynamic right-sizing algorithm can achieve significant energy savings in the data center. We should consider that this requires servers to have different power modes and be able to transit to different states while still keeping the previous state. Moving the system to different power consumption modes is a challenging problem and requires dynamic on-line policies for resource management.

Green Open Cloud (GOC) is an architecture which is proposed by Lefèvre and Orgerie (2010) on top of the current resource management strategies. The aim of this architecture is to switch-off unused servers, predict the incoming requests, and then switch-on required servers on the arrival of new requests. GOC proposes green policies to customers in the way that they can have advance resource reservation and based on this knowledge cloud provider could estimate how many servers, and when they should be switched-on. Using this framework and strategy, they were able to save a considerable amount of energy on cloud severs.

Prediction-Based Algorithms

Aksanli, Venkatesh, Zhang, and Rosing (2012) used the data from solar and wind power installations in San Diego (MYPVDATA) and National Renewable Energy Laboratory (NREL), respectively to develop a prediction-based scheduling algorithm to serve two different types of workloads, web-services and batch-jobs. The main goal of this model was to increase the efficiency of the green energy usage in data centers. Based on the experiments of the proposed model, the number of tasks that were done by the green energy resources increased and the number of works that were terminated because of the lack of enough green energy resources decreased. This model uses a single queue per server for web services which are time sensitive applications, and for the batch-jobs it uses the Hadoop which is the general form of Map-Reduce framework.

GreenSlot scheduler (Goiri t al., 2011) also proposes a scheduling and prediction mechanism to efficiently use the green energy sources. Goiri et al. (2011) consider solar as the main source of energy and smart grid, known as brown energy, as the backup power source for the data center. The main objective of GreenSlot is to predict the availability of solar energy two days in advance so that it can maximize the use of green energy and reduce the costs associated with using brown energy. GreenSlot uses the suspension mechanism when there is not enough green energy available and based on the availability of enough solar energy it resumes the jobs. According to the experimental results that are presented in comparison with other conventional scheduling mechanisms, like backfilling scheduler (Mu'alem & Feitelson, 2001), GreenSlot scheduler can significantly increase the use of green energy for running batch-jobs and decrease the brown energy costs, which leads to less carbon footprint and moving towards a sustainable environment. Unlike web-service jobs which are time-sensitive batch-jobs are compute intensive and the deadline is not critical as web-service jobs, so the suspension will not affect the user quality of service (QoS) parameters.

VM Placement

Users send their requests to the cloud Infrastructure as a Service (IaaS) providers in the form of VMs. Goudarzi and Pedram (2012) presented a VM placement heuristic algorithm to place the VMs in physical servers in a way to reduce data centers energy consumption. The algorithm receives the VM requests and splits each VM into several copies and places them on servers. Each copy of VM gets the same amount of physical memory but with different CPUs. The total summation of assigned CPUs for copies of a VM request will be equal to the required CPU by the VM at the time of arrival to the data center. The proposed algorithm, which is known as MERA (Multi-dimensional Energy-efficient Resource Allocation), receives the VM requests and after a certain time epoch places the VMs on the servers and calculates the consumed energy. Then, it splits the VMs and places the copies on servers and recalculates the energy consumption. Based on the calculated energies the algorithm makes decision whether to split and replicate VMs or not. This algorithm tries to increase the servers' utilization while decreasing the energy consumption without considering the physical characteristics and energy related parameters of servers and data centers. Moreover, it does not perform the VM placement dynamically. The algorithm receives a group of VMs and after a certain time epoch performs VM placement. In addition, intercommunication between replicated VMs could lead to bottleneck and high energy consumption. Finally, in the placement all VMs are treated the same. As all the replicated VMs get the same amount of physical memory, whilst for memory-intensive VMs this could result to shortage in resources and it is better to make balance between CPU intensive and memory intensive VM requests.

The work done by Xu and Fortes (2010) addresses the problem of data centers VM placement with the objective to simultaneously minimize resource wastage, power consumption, and maximum temperature of the servers. They used a genetic algorithm on the global controller of the data center to perform the VM placement. The global controller receives the VM requests and then based on a multi-objective VM placement algorithm assigns each VM to a server. This algorithm, same as the previously discussed work, performs VM placement after receiving all the VM requests, which is not in a dynamic manner. Moreover, the algorithm makes balance between power consumption and temperature. Therefore, it uses more servers to distribute the load and avoid hotspots in the data center. This might cause more carbon footprint as more servers will be used and more electricity will be consumed.

Green SLA Aware

Due to the high energy consumption by cloud data centers and climate concerns, cloud providers do not just rely on the electricity coming from brown energy sources. They have their own on-site green energy sources or draw it from a nearby power plant. Moreover, enterprises and individuals demand for quantifiable green cloud services. Haque, Le, Goiri, Bianchini, and Nguyen (2013) propose a new class of cloud services that provides a specific service level agreement for users to meet the required percentage of green energy used to run their workloads. They undertake a new power infrastructure in which each rack can be powered from brown or green energy sources. The optimization policies have the objective of increasing the provider's profit by admitting the incoming jobs, with Green SLA requirements. If cloud provider cannot meet the requested percentage of green energy to run the job should pay penalty to the user, which means decrease in the total gained profit of running jobs. The type of green energy used by Haque et al. (2013) in the data center is solar energy and they predict the availability and amount of solar energy based on the method proposed in Goiri et al. (2012). The experiments carried in their

work are based on comparison with greedy heuristics, and they show that optimization based policies outperform the greedy ones. Furthermore, among optimization based policies cloud provider can decide whether wants to increase the number of admitted jobs or violate less Green SLAs.

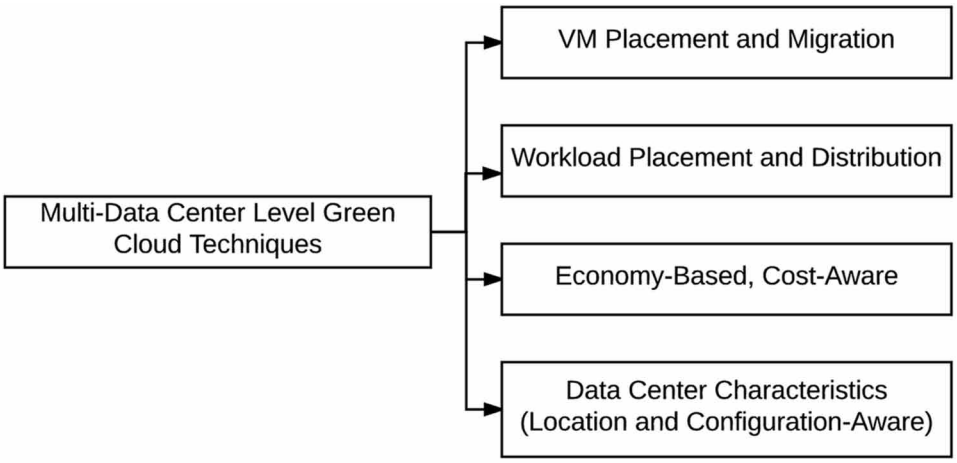
In the calculated total cost to run the admitted jobs in the work by Haque et al. (2013), it is not clear that whether it is the cost to run the servers or the total cost in the data center, including overhead energy cost as well. This is important because overhead energy is dependent on the data center power usage effectiveness (PUE) and this varies by the change in the data center total utilization and ambient temperature (Rasmussen, 2007; Goiri, Le, Guitart, Torres, & Bianchini, 2011). Therefore, the calculated value for profit in the optimization based policies would vary based on the two aforementioned parameters for different jobs with different configuration requirements and also time of the day.

ENERGY EFFICIENCY IN GEOGRAPHICALLY DISTRIBUTED DATA CENTERS

Applying different policies to switch-off and on servers and placing user requests within a data center could lead to reduce in energy consumption. But still these are not enough to solve the problem of high energy consumption and carbon footprint by cloud data centers.

By increasing the use of cloud computing services that leads to increase in energy consumption and carbon footprint in the environment, some cloud providers decided to use green energy as a secondary power plant. Therefore, the need to have a scheduling policy to select the data center site to run the user request based on the energy source is necessary. Moreover, data center selection based on considering different data centers energy efficiency, as it has a direct effect on total carbon footprint, reduces energy consumption and carbon dioxide in the ecosystem. This section explores different energy and carbon-efficient approaches have been taken across distributed cloud data centers. Some of the applied techniques are the same as single data center level, but with considering factors to select the data center site before cluster and server selection. Figure 4 shows the taxonomy of different approaches taken at multi data center level with different optimization objectives, such as minimizing cost, energy consumption, carbon emission, and maximizing renewable energy consumption.

Figure 4. Multi data center level energy and carbon-efficient techniques



VM Placement and Migration

Research works in this area consider initial placement of a VM and further monitoring of the running VM to meet the optimization objective. Virtual machine (VM) placement in a geographically distributed data center environment requires selection of a data center and a server within the data center based on the optimization objective and data centers characteristics. Moreover, after the VM placement considering the future state of the host data center and other data centers, cloud provider can perform VM live migration to move the VM to another data center with preferable parameters. There are a few research works that consider these two techniques.

Chen, Xu, Xi, and Chao (2013) developed a model for optimal VM placement considering a cloud provider with distributed data center sites connected through leased/dedicated lines. They introduce a cost-aware VM placement problem with the objective of reducing operational cost as a function of electricity costs to run the VMs and inter-data center communication costs. For this purpose, they take advantage of variable electricity costs at multiple locations and wide-area network (WAN) communication cost to place the VMs using a meta-heuristic algorithm. Similarly, Qureshi, Weber, Balakrishnan, Guttag, and Maggs (2009) try to minimize electricity cost of running the VMs by initially placing the VMs into data centers with low spot market prices. They take advantage of spatial and temporal variations of electricity price at different locations.

Akoush, Sohan, Rice, Moore, and Hopper (2011) propose an architecture known as Free Lunch to maximize renewable energy consumption. They consider having data center sites in different geographical locations in such a way to complement each other in terms of access to renewable energy (solar and wind) by being located in different hemisphere and time zone. The architecture considers pausing VMs execution in the absence of renewable energy or migrating the VMs to another data center site with excess renewable energy. The proposed architecture provides a good insight to harness renewable energy by having geo-distributed data center sites with dedicated network. However, this model has technical challenges and limitations dealing with VM availability, storage synchronization, VM placement and migration that have been pointed out in their work.

Work done by Khosravi, Garg, and Buyya (2013) addresses the problem of energy consumption and carbon footprint of distributed cloud data centers by proposing a novel framework and algorithm for VM placement. This system model uses a component known as Cloud Information Service (CIS) in order to get the data centers' information and updates to perform the scheduling algorithm. The information a data center sends to the CIS consists of data center's available resources, energy and carbon related parameters, such as power usage effectiveness, carbon footprint rate/s (a data center might use more than one energy source), and servers' proportional power as a metric related to the CPU frequency and utilization. The cloud broker, as the interface between users and cloud provider, uses this information to perform a dynamic two-level scheduling algorithm. The algorithm places the VM in the data center/cluster with the least carbon footprint and energy consumption (first level), and in the server with the least increase in the power consumption (second-level), while meeting the users' quality of service in terms of number of rejected VMs. The proposed algorithm reduces the carbon footprint and energy consumption considerable in comparison to other competing algorithms.

Workload Placement and Distribution

A large body of literature recently focused on reducing energy consumption and energy costs by load placement and distribution across geographically distributed data centers.

Le, Bianchini, Martonosi, and Nguyen (2009) proposed a framework to reduce cost and brown energy consumption of cloud computing systems by distributing user requests across data center sites. This is the first research that considers load distribution across data center sites with respect to their energy source and cost. The framework is composed of a front-end that receives user requests and based on a distribution policy forwards the requests to the data center site with less cost and more available green energy sources. The request distribution policy sorts the data center sites based on the percent of the load that could be completed within a time period and minimum cost to run the requests. The evaluation results show that by knowing data centers' electricity price (constant price, dynamic, or on/off-peak prices) and base/idle energy consumption of the servers', significant improvements in cost reduction will be made. Moreover, being aware of the energy sources (green or brown) in the data centers could lead to less brown energy usage with a slight increase in the total cost.

Zhang, Wang, and Wang (2011) use the idea of distributing the load among a network of geographically distributed data centers to maximize renewable energy usage. They proposed a novel middleware, known as GreenWare, that dynamically conducts user requests to a network of data centers with the objective of maximizing the percentage of renewable energy usage, subject to the cloud service provider cost budget. Experiment results show GreenWare could significantly increase the usage of renewable energies, solar and wind with intermittent nature, whilst still meeting the cost budget limitation of the cloud provider.

Following the idea of reducing brown energy consumption in data center sites, Liu, Lin, Wierman, Low, and Andrew (2011) proposed the geographical load balancing (GLB) algorithm. The algorithm takes advantage of diversity of data center sites to route requests to the places with access to renewable, solar and wind, energy sources. Considering the unpredictable nature of renewable energy, specially wind, GLB algorithm finds the optimal percentage of wind/solar energies to reduce the brown energy consumption and carbon footprint. Moreover, the authors consider the role of storage of renewable energies, when they are not available in data centers in reducing brown energy usage. Based on the experiments, by using even small-scale storage in the data centers, the need for brown energy will decrease and in some cases even will be eliminated. A question that might rise with Liu et al. (2011) work is the carbon footprint caused by the batteries in a long-term period, since renewable energy storage in the data center sites is done through reserving them in the form of batteries. Lin, Liu, Wierman, and Andrew (2012) extended the GLB algorithm to reduce the total cost along with reducing the total brown energy consumption for geographically distributed data centers. They compared their proposed algorithm with two prediction-based algorithms with a look-ahead window, known as receding horizon control (RHC) a classical control policy and an extension of RHC known as averaging fixed horizon control (AFHC) (Kwon & Pearson, 1977). The analytical modelling and the simulations carried, based on real workload traces, show that GLB algorithm can reduce the energy cost by slightly increase in network delay. Moreover, it can eliminate the use of brown energy sources by routing user requests to the sites where wind/solar energy is available.

Garg, Yeo, Anandasivam, and Buyya (2011) proposed an environment-conscious meta-scheduler for high performance computing (HPC) applications in a distributed cloud data center system. The meta-scheduler consists of two phases, mapping the applications to the data center and scheduling within

a data center. They treat the mapping and scheduling of applications as an NP-hard problem with the objective to reduce carbon emission and increase the cloud provider profit at the same time. They run different experiments in order to find the near optimal solution for this dual objective problem. The parameters taken into account in the simulations and scheduling algorithms are data centers' carbon footprint rate, electricity price, and data center's efficiency. The simulations carried for high urgent applications (with short deadlines) and different job arrival rates help the cloud providers to decide for each application which scheduling algorithms should be used in order to meet the objective of reducing the carbon emission or maximizing the profit. Moreover, they proposed a lower bound and an upper bound for the carbon emission and profit, respectively. Another work done by Garg, Yeo, and Buyya (2011) addresses the issue of energy efficiency of ICT industry, specially data centers. The main focus of this work is to reduce the carbon footprint of running workloads on data centers by proposing a novel carbon-aware green cloud architecture. This architecture consists of two directories, which imposes the use of green energy by data centers while meeting users and providers' requirements. In this framework, cloud providers should register their offered services in the aforementioned directories, and the users should submit their requests to the data centers through the Green Broker. The scheduling mechanism used in the broker, Carbon Efficient Green Policy (CEGP), chooses the cloud provider based on the least carbon footprint while considering users QoS parameters. The performance evaluation results of the proposed framework and policy in comparison with a traditional scheduling approach shows that CEGP can achieve a considerable reduction in energy consumption and carbon footprint in the ecosystem. However, this algorithm does not work dynamically. It receives all the job requests and based on the jobs deadline assigns them to the data center with the least carbon footprint. Moreover, it only considers high performance computing applications (non-interactive workloads) with predefined deadlines at the time of submission.

Chen, He, and Tang (2012) use the idea of geographically distributed data centers to increase usage of green energy and reduce brown energy consumption in data centers. They proposed a workload scheduling algorithm, called MinBrown, that considers green energy availability in different data centers with different time zones, cooling energy consumption for data centers based on outside temperature and data center utilization, incoming workload changes during time, and deadline of the jobs. The workload used to run the simulation is HPC jobs with sufficient slack time to allow advanced scheduling. The algorithm copies all the data in all the data centers and based on the least consumed brown energy executes the task. Based on the simulation results, the MinBrown algorithm reduces brown energy consumption in comparison to other competitive algorithms. The idea of replicated data in distributed data center sites itself results to high energy consumption that is not considered in Chen et al. (2012) work. Moreover, assignment of the jobs and tasks are based on the availability of green energy, that does not consider communication between tasks of the same job and jobs of the same workload. Finally, the scheduler does not consider an efficient resource assignment within a data center in a way to reduce the need for future consolidation of the running jobs.

The idea of federation of cloud providers can be useful for relocation of computational workload among different providers in a way to increase the use of sustainable energy. Celesti, Puliafito, Tusa, and Villari (2013) take advantage of a federated cloud scenario to reduce energy costs and CO_2 emissions. They consider cloud providers' data centers are partially powered by renewable energies along with getting the required electrical energy from electrical grids. The main contribution of their work is based on the approach of moving the workload towards the cloud data center with most available sustainable energy. This is inspired by the fact that if a provider generates more green energy than its

need, it would be difficult to store the exceeded amount in batteries or put it in public grids; therefore, the easiest way is to relocate the workload to the site with the excess renewable energy. The architecture is based on an Energy Manager, that is known as CLoud-Enabled Virtual Environment (CLEVER). By applying CLEVER-based scenario, the VM allocation would be based on the energy and temperature driven policies. The energy manager in the architecture receives different data centers' information, such as temperature, sun radiation, energy grid fare, photovoltaic energy, cost, and data centers' PUE and number of available slots or physical resources, and based on this data assigns VMs to the site with the most sustainable energy and least cost.

Celesti et al. (2013) work increases the use of sustainable energies and it is based on the availability of the photovoltaic (PV) energy. When a site has a high value for the PV energy, the outside temperature would be higher and this will increase the need for more energy for the cooling, and as a result higher PUE value. Relying only on the amount of used PV in the system is not enough for a green and sustainable system. Cloud providers should consider the whole picture and take into account all the parameters that affect the total CO_2 emission. Moreover, Celesti et al. (2013) assume that each new VM request would be replicated in all the federated providers. Considering the consumed energy for this replication and the effect of network distance are also important that should be considered by the time of system design.

Xu, Feng, and Li (2013) take advantage of diversity in data centers location to route the incoming workload with the objective of reducing the energy consumption and cost. They studied the effect of ambient temperature on the total energy consumed by cooling system, which is 30% to 50% of the total energy consumption of data centers (Pelley, Meisner, Wenisch, & VanGilder, 2009; Zhou et al., 2012). Energy consumption often is modelled as a constant factor, which is an over-simplification of what is happening in reality. Xu et al. (2013) considered partial PUE (power usage effectiveness) to participate cooling systems' energy along with the servers' total energy consumption. Through using partial PUE data centers can route the workload to the sites that use outside air cooling and reduce considerable amount of energy consumption. Moreover, they took advantage of having two types of incoming requests to manage the resources and reduce the energy consumption. The proposed model does not only depend on the energy consumed by interactive workload form users, instead it reduces energy costs by allocating capacity to the batch workloads, which are delay tolerant and can be run at the back-end of the data centers. The proposed joint optimization approach could reduce cooling energy and overall energy cost of data centers.

However, the proposed partial PUE only considers the energy consumed by cooling system as the total overhead energy in the data center. Based on the introduced definition by Xu et al. (2013), PUE is mainly dependent on the ambient temperature, while IT load of the data center is the second important factor affecting the PUE (Rasmussen, 2007). Finally, source of the energy used to generate the electricity and its carbon footprint is not considered. This is important because as mentioned earlier reducing energy cost does not necessarily lead to reduce in the carbon footprint in the environment.

Economy-Based, Cost-Aware

Cost associated with energy usage in large data centers is a major concern for the cloud providers. Large data centers consume megawatts of electricity, which leads to huge operational costs. Work done by Ren, He, and Xu (2012) takes advantage of different electricity prices in different geographical locations and over time to schedule batch jobs on the servers in scattered data centers. Their proposed online optimal algorithm, known as GreFar, uses servers' energy efficiency information and locations with low elec-

tricity prices to schedule the arrived batch jobs from different organizations. GreFar's key objective is to reduce energy cost, while assuring fairness considerations and delay constraints. The scheduling is based on a provably-efficient online algorithm, that schedules the jobs according to the current job queue lengths. Based on the simulation results, GreFar online algorithm can reduce system cost, in terms of a combination of energy cost and fairness, in comparison to the offline algorithm that has knowledge of system's future state. The algorithm's main contribution is to serve the jobs when the electricity price is low or there are energy-efficient servers in the system. To accomplish this objective, it queues jobs and suspends low priority jobs whenever the electricity price goes up or there are not enough efficient servers in the system. This approach is not applicable for interactive jobs and web requests that are time sensitive and need to be served immediately from the queue and also cannot be suspended. Moreover, the cloud provider does not consider the cost of the transmission network and its energy consumption at the time of data center selection to submit the job request.

Le, Bianchini, Nguyen, Bilgir, and Martonosi (2010) take advantage of capping the brown energy consumption to reduce the cost of serving Internet services in data centers. They proposed an optimization-based framework to distribute requests among distributed data centers, with the objective to reduce costs, while meeting users' service level agreement (SLA). The main parameters that affect the site selection by the framework are different electricity prices (on-peak and off-peak loads), different data centers location with different time zones, data centers with access to green energy sources, which enables the data center to have a mixture of brown and green energy. The front-end of the framework performs the site selection and optimization problem for the arrived requests periodically, in contrast to heuristic algorithms, which are greedy and select the best destination for each request that arrives (Qureshi et al., 2009). The optimization framework uses load prediction by Auto-Regressive Integrated Moving Average (ARIMA) modeling (Box, Jenkins, Reinsel, & Ljung, 2015) and simulated annealing (SA) (Brooks & Morgan, 1995) to predict the load for the next epoch (one week) and schedule the requests. This approach helps the front-end to decide about the power mixes at each data center for the next week, unless a significant change occurs in the system and predictions. Le et al. (2010) use simulation and real system experiments with real traces to evaluate their proposed framework and optimization policy. The evaluation results show that by taking optimization policy and using workload prediction, diversity in electricity price, taking benefit of brown energy caps, and use of green energy sources significant savings in cost related to the execution of Internet services in distributed data centers would be made. The framework assumes that all the received requests from the users are homogeneous. While in the real systems this is not the case and having heterogeneous requests and distributing them in a way to reduce resource wastage is very difficult and itself results to huge energy consumption and accordingly high costs. Moreover, it focuses on the electricity prices in different locations without considering the carbon footprint rate of the sources. Since some brown energy sources, which are cheap and lead to reducing the system overall cost, may lead to huge amount of carbon dioxide in the ecosystem.

The other work by Le et al. (2011) investigates different parameters that affect the electricity costs for geographically distributed data centers with the focus on IaaS services that run HPC workloads. According to their proposed cost computation framework for the data centers, there are two important parameters that affect the total cost, energy consumed to run the service and the cost for the peak power demand. The provider can reduce the consumed energy by selecting the sites with off-peak period electricity prices, lower outside temperature, and lower data center load, so that the energy used for cooling would be low. Because as the data center temperature rises, the provider needs to use chillers to reduce temperature, which increases the energy consumption dramatically. In order to show this relation, they

used a simulation model for the data center cooling system. Based on the simulation model, increase in the outside temperature and data center load forces the providers to use the chillers in order to keep the data center cool. This simulation has been carried with real workload traces from the Feitelson (2007), Parallel Workloads Archive. Le et al. (2011) compared their two proposed algorithms, cost-aware and cost-aware with migration, with baseline policies. Based on the results, considering above mentioned factors can reduce the energy cost of data centers. Moreover, predicting the need to use the chillers for system cooling and considering the transient cooling prevents the data center from overheating and would not let spikes in the temperature.

Le et al. (2011) conducted sensitivity analysis to investigate the effect of parameters, such as predicting the run-time of the jobs, the time to migrate the jobs, outside temperature, price of the energy in a region, and size of the data center on the total cost of the data center. According to the simulation results, in order to maximize the cost-saving all the electricity-related parameters should be considered in job placement in the system. One of the shortcomings of this work, similar to the previously discussed work, is not considering the source of electricity. As some brown energy sources with high carbon footprint might be cheaper and more desirable to run the services. Moreover, as the temperature changes during the day and the consumed energy for cooling changes consequently; PUE should be modelled as a dynamic parameter instead of having a constant value per data center. Considering network distance and the energy consumption of intra and inter-data centers will also affect the total cost.

Work by Buchbinder, Jain, and Menache (2011) has also the objective of reducing energy cost for a cloud provider with multi data center sites but with a different approach. They perform on-line migration of running batch jobs among data center sites, taking advantage of dynamic energy pricing and power availability at different locations, while considering the network bandwidth costs among data centers and future changes in electricity price. The total cost in their model, is the cost of energy to run the jobs at the destined data center plus the bandwidth cost to migrate the data. To attain an optimal algorithm with lower complexity comparing the optimal off-line solution, Buchbinder et al. (2011) proposed an efficient on-line algorithm (EOA) with higher performance comparing to the greedy heuristics that ignore the future outcomes. The calculated cost in their work is based on the data centers' operational cost, which focuses on the energy consumption by servers and transport network. However, a considerable part of the energy consumed by a data center is related to the overhead energy, such as cooling systems. Moreover, the objective of reducing the energy cost and routing the jobs to the data center with lowest cost without considering the energy source might lead to increase in the carbon footprint in the environment. The migration of running jobs in this work is in the context of batch jobs, which are delay tolerant in comparison to user interactive requests such as web requests that are delay sensitive. Therefore, the applicability of this algorithm should be investigated for other workloads and user requests in a cloud computing environment. Similarly, work by Luo, Rao, and Liu (2015) leverages both the spatial and temporal variation of electricity price to route the incoming requests between geographically distributed data centers targeting energy cost minimization.

Data Center Characteristics (Location and Configuration-Aware)

There are several works try to make data centers energy and carbon-efficient by reducing the number of active servers or run the virtual machines and applications on the physical machines with the least energy consumption and carbon footprint rate. However, geographical location of the data center has a direct impact on the amount of consumed energy that leads to CO_2 emission in the ecosystem. Work done by

Goiri et al. (2011) considers intelligent placement of data centers for Internet services. Their goal is to find the best location for data center site to minimize the overall cost and respect users' response time, consistency, and availability. They classified the parameters that affect data centers overall cost into location dependent and data center characteristics data.

The location dependent data specifies the data center's distance to the network backbones, power plants, and the CO_2 emission of the power plant. Moreover, it includes the electricity, land, and water price. The last and one of the most important factors related to the location is the outside temperature. Since, when the temperature goes high the need for cooling increases as well. Cooling system is an important parameter in the data centers, which its energy consumption increases as outside temperature increases. Indeed, high temperature leads to need for more chillers and more chillers increases data center's total energy consumption. This situation eventually leads to higher PUE and energy consumption, which indirectly increases carbon footprint. Goiri et al. (2011) propose a framework to find the most optimum location for the data center to minimize the total costs. Explicit decrease in data center's cost, leads to indirect decrease in energy consumption and carbon footprint.

In order to increase the use of renewable energies, Berral et al. (2014) propose a framework to find the best location to site the data centers and renewable power plants, solar and wind in their work. In the meantime, their objective is reducing total cost for building these infrastructures to support cloud HPC services with different amounts of renewable usage. Berral et al. (2014) divided the costs of building green cloud services into capital (CAPEX) and operational (OPEX) costs and CAPEX itself is divided to costs dependent and the costs that are independent to the number of servers to be hosted. Independent CAPEX costs are cost of bringing brown energy to the data center and connecting to the backbone network. Land cost, building green power plants, cooling infrastructure, batteries, networking equipment, and servers are part of the dependent CAPEX costs. Costs incurred during the life cycle of the data center, such as network bandwidth and amount of brown energy usage are part of the OPEX. Brown energy consumption is the total energy needed by the servers and overhead parts, such as cooling and networking, minus energy derived from renewables. To calculate the overhead energy, Berral et al. (2014) use PUE as a parameter related to the location temperature. It should be noted that temperature is not the only parameter that affects PUE, data center load is also an important parameter that changes PUE value (Rasmussen, 2007).

In order to take the most of the generated renewable energy in different data centers, Berral et al. (2014) compare different approaches such as net metering, which is directing the excess renewable energy into the grid and mix it with brown energy, using batteries and having storage for renewables or not having any storage and migrating the load to the sites with available solar or wind. One of the shortcomings of their work is neglecting the network delay and amount of energy consumed due to VM migration, as the

data centers are scattered at different geographical locations. Moreover, all the data in this system are replicated at all the sites, which itself imposes overhead and increases energy consumption.

CONCLUSION AND FUTURE DIRECTIONS

In this chapter, the authors studied the research works in the area of energy and carbon footprint-aware resource management in cloud data centers. They first had an overview on the existing techniques in green cloud resource management with the focus on a single server and a single data center and the limitations facing these techniques, specially not being able to harvest renewable energy sources at different locations.

The authors then focused more specifically on the works considering geo-distributed cloud data centers, as nowadays most of the big cloud providers have data centers in different geographical locations for disaster recovery management, higher availability, and providing better quality of experience to users.

A large body of literature in the context of distributed data centers considers assigning resources to the arrived requests in such a way to minimize brown energy consumption. They use different techniques such as applying VM placement heuristics, workload scheduling, and targeting data centers with the most available renewable energy. These works explicitly consider access to renewable energy sources to minimize brown energy consumption and carbon footprint. However, some of the research works achieve energy efficient resource management through minimizing cost and the cost of brown energy usage, which indirectly could lead to less carbon footprint in the ecosystem.

Research in the area of energy and carbon-efficient resource management in data centers is still an important field of work. Apart from the surveyed techniques in this chapter, there are still areas that can be pursued by researchers. VM migration across data center sites to harvest the renewable energy sources is still at its early stages. First of all, it is important to study the effect of minimizing brown energy usage and carbon cost versus network cost and delay imposed due the data transfer over the network. Selecting the VMs to migrate depending on the application running on top of the VM with respect to users' service level agreement is also another area of future study.

There are studies that consider storing excess renewable energy in batteries to use at times of the day that renewable sources are not available. Since main cloud providers started to build their own on-site renewable energy sources and having large scale renewable energy power plants, studying the cost-effectiveness of storing the renewable energy for future usage and contributing to the electrical grid is an important area for future study.

REFERENCES

Akoush, S., Sohan, R., Rice, A. C., Moore, A. W., & Hopper, A. (2011). Free Lunch: Exploiting Renewable Energy for Computing. HotOS, 13, 17.

Aksanli, B., Venkatesh, J., Zhang, L., & Rosing, T. (2012). Utilizing green energy prediction to schedule mixed batch and service jobs in data centers. *SIGOPS Operating Systems Review, ACM*, *45*(3), 53–57. doi:10.1145/2094091.2094105

Baer, P. (2008). *Exploring the 2020 global emissions mitigation gap. Analysis for the Global Climate Network*. Stanford University, Woods Institute for the Environment.

Barroso, L. A., & Holzle, U. (2007). The case for energy-proportional computing. *IEEE Computer*, *40*(12), 33–37. doi:10.1109/MC.2007.443

Beloglazov, A., Buyya, R., Lee, Y. C., & Zomaya, A. (2011). A taxonomy and survey of energy-efficient data centers and cloud computing systems. *Advances in Computers, 82*(2), 47-111.

Berral, J. L., Goiri, Í., Nguyen, T. D., Gavalda, R., Torres, J., & Bianchini, R. (2014). Building green cloud services at low cost.*34th International Conference on Distributed Computing Systems (ICDCS)*, 449-460. doi:10.1109/ICDCS.2014.53

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brey, T., & Lamers, L. (2009). Using virtualization to improve data center efficiency. *The Green Grid, Whitepaper, 19*.

Brooks, S. P., & Morgan, B. J. (1995). Optimization using simulated annealing. *The Statistician*, *44*(2), 241–257. doi:10.2307/2348448

Buchbinder, N., Jain, N., & Menache, I. (2011). Online job-migration for reducing the electricity bill in the cloud.*International Conference on Research in Networking*, 172-185. doi:10.1007/978-3-642-20757-0_14

Celesti, A., Puliafito, A., Tusa, F., & Villari, M. (2013). Energy Sustainability in Cooperating Clouds. CLOSER, 83-89.

Chen, C., He, B., & Tang, X. (2012). Green-aware workload scheduling in geographically distributed data centers.*4th International Conference on Cloud Computing Technology and Science (CloudCom)*, 82-89. doi:10.1109/CloudCom.2012.6427545

Chen, K. Y., Xu, Y., Xi, K., & Chao, H. J. (2013). Intelligent virtual machine placement for cost efficiency in geo-distributed cloud systems.*International Conference on Communications (ICC)*, 3498-3503. doi:10.1109/ICC.2013.6655092

Feitelson, D. (2007). *Parallel workloads archive*. Academic Press.

Garg, S. K., Yeo, C. S., Anandasivam, A., & Buyya, R. (2011). Environment-conscious scheduling of HPC applications on distributed Cloud-oriented data centers. *Journal of Parallel and Distributed Computing*, *71*(6), 732–749. doi:10.1016/j.jpdc.2010.04.004

Garg, S. K., Yeo, C. S., & Buyya, R. (2011). Green cloud framework for improving carbon efficiency of clouds.*European Conference on Parallel Processing*, 491-502. doi:10.1007/978-3-642-23400-2_45

Goiri, I., Le, K., Guitart, J., Torres, J., & Bianchini, R. (2011). Intelligent placement of datacenters for internet services.*31st International Conference on Distributed Computing Systems (ICDCS)*, 131-142. doi:10.1109/ICDCS.2011.19

Goiri, Í., Le, K., Haque, M. E., Beauchea, R., Nguyen, T. D., Guitart, J., & Bianchini, R. (2011). GreenSlot: scheduling energy consumption in green datacenters.*Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*, 20. doi:10.1145/2063384.2063411

Goiri, Í., Le, K., Nguyen, T. D., Guitart, J., Torres, J., & Bianchini, R. (2012). GreenHadoop: leveraging green energy in data-processing frameworks.*Proceedings of the 7th ACM european conference on Computer Systems*, 57-70. doi:10.1145/2168836.2168843

Goudarzi, H., & Pedram, M. (2012). Energy-efficient virtual machine replication and placement in a cloud computing system.*5th International Conference on Cloud Computing (CLOUD)*, 750-757. doi:10.1109/CLOUD.2012.107

Haque, M. E., Le, K., Goiri, Í., Bianchini, R., & Nguyen, T. D. (2013). Providing green SLAs in high performance computing clouds.*International Green Computing Conference (IGCC)*, 1-11. doi:10.1109/IGCC.2013.6604503

Harney, E., Goasguen, S., Martin, J., Murphy, M., & Westall, M. (2007). The efficacy of live virtual machine migrations over the internet.*Proceedings of the 2nd international workshop on Virtualization technology in distributed computing*, 8. doi:10.1145/1408654.1408662

Khosravi, A., Garg, S. K., & Buyya, R. (2013). Energy and carbon-efficient placement of virtual machines in distributed cloud data centers.*European Conference on Parallel Processing*, 317-328. doi:10.1007/978-3-642-40047-6_33

Koomey, J. G. (2007). *Estimating total power consumption by servers in the US and the world*. Academic Press.

Koomey, J. G. (2008). Worldwide electricity used in data centers. *Environmental Research Letters*. *IOP Publishing*, *3*(3), 034008.

Kwon, W., & Pearson, A. (1977). A modified quadratic cost problem and feedback stabilization of a linear system. *IEEE Transactions on Automatic Control*, *22*(5), 838–842. doi:10.1109/TAC.1977.1101619

Le, K., Bianchini, R., Martonosi, M., & Nguyen, T. D. (2009). Cost-and energy-aware load distribution across data centers. Proceedings of HotPower, 1-5.

Le, K., Bianchini, R., Nguyen, T. D., Bilgir, O., & Martonosi, M. (2010). Capping the brown energy consumption of internet services at low cost.*International Green Computing Conference*, 3-14. doi:10.1109/GREENCOMP.2010.5598305

Le, K., Bianchini, R., Zhang, J., Jaluria, Y., Meng, J., & Nguyen, T. D. (2011). Reducing electricity cost through virtual machine placement in high performance computing clouds.*Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*, 22. doi:10.1145/2063384.2063413

Lefèvre, L., & Orgerie, A. C. (2010). Designing and evaluating an energy efficient cloud. *The Journal of Supercomputing*, *51*(3), 352–373. doi:10.1007/s11227-010-0414-2

Lin, M., Liu, Z., Wierman, A., & Andrew, L. L. (2012). Online algorithms for geographical load balancing.*International Green Computing Conference (IGCC)*, 1-10.

Lin, M., Wierman, A., Andrew, L. L., & Thereska, E. (2013). Dynamic right-sizing for power-proportional data centers. *Transactions on Networking (TON), IEEE/ACM, 21*(5), 1378-1391.

Liu, Z., Lin, M., Wierman, A., Low, S. H., & Andrew, L. L. (2011). Geographical load balancing with renewables. *Performance Evaluation Review*, *39*(3), 62–66. doi:10.1145/2160803.2160862

Luo, J., Rao, L., & Liu, X. (2015). Spatio-Temporal Load Balancing for Energy Cost Optimization in Distributed Internet Data Centers. *IEEE Transactions on Cloud Computing*, *3*(3), 387–397. doi:10.1109/TCC.2015.2415798

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. NIST special publication, 800, 145.

Mualem, A. W., & Feitelson, D. G. (2001). Utilization, predictability, workloads, and user runtime estimates in scheduling the IBM SP2 with backfilling. *IEEE Transactions on Parallel and Distributed Systems*, *12*(6), 529–543. doi:10.1109/71.932708

MYPVDATA Energy Recommerce. (n.d.). Retrieved from https://www.mypvdata.com/

National Renewable Energy Laboratory (NREL). (n.d.). Retrieved from http://www.nrel.gov/

Pelley, S., Meisner, D., Wenisch, T. F., & VanGilder, J. W. (2009). Understanding and abstracting total data center power.*Workshop on Energy-Efficient Design*.

Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J., & Maggs, B. (2009). Cutting the electric bill for internet-scale systems. ACM SIGCOMM Computer Communication Review, 39(4), 123-134. doi:10.1145/1592568.1592584

Rasmussen, N. (2007). Electrical efficiency measurement for data centers. *White paper, 154*.

Ren, S., He, Y., & Xu, F. (2012). Provably-efficient job scheduling for energy and fairness in geographically distributed data centers.*32nd International Conference on Distributed Computing Systems (ICDCS)*, 22-31. doi:10.1109/ICDCS.2012.77

Srikantaiah, S., Kansal, A., & Zhao, F. (2008). Energy aware consolidation for cloud computing.*Proceedings of the conference on Power aware computing and systems*, 10, 1-5.

Xu, H., Feng, C., & Li, B. (2013). Temperature aware workload management in geo-distributed datacenters.*Proceedings of the 10th International Conference on Autonomic Computing (ICAC)*, 303-314. doi:10.1145/2465529.2465539

Xu, J., & Fortes, J. A. (2010). Multi-objective virtual machine placement in virtualized data center environments.*Int'l Conference on Green Computing and Communications (GreenCom) & Int'l Conference on Cyber, Physical and Social Computing (CPSCom), IEEE/ACM*, 179-188. doi:10.1109/GreenCom-CPSCom.2010.137

Zhang, Y., Wang, Y., & Wang, X. (2011). Greenware: Greening cloud-scale data centers to maximize the use of renewable energy.*International Conference on Distributed Systems Platforms and Open Distributed Processing, ACM/IFIP/USENIX*, 143-164. doi:10.1007/978-3-642-25821-3_8

Zhou, R., Wang, Z., McReynolds, A., Bash, C. E., Christian, T. W., & Shih, R. (2012). Optimization and control of cooling microgrids for data centers.*13th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 338-343. doi:10.1109/ITHERM.2012.6231449