Harnessing Renewable Energy in Cloud Datacenters: Opportunities and Challenges

Wei Deng, Fangming Liu, and Hai Jin, Services Computing Technology and System Lab, Cluster and Grid Computing Lab in the School of Computer Science and Technology, Huazhong University of Science and Technology

Bo Li, The Hong Kong University of Science and Technology

Dan Li, Tsinghua University

Abstract

The proliferation of cloud computing has promoted the wide deployment of large-scale datacenters with tremendous power consumption and high carbon emission. To reduce power cost and carbon footprint, an increasing number of cloud service providers have considered green datacenters with renewable energy sources, such as solar or wind. However, unlike the stable supply of grid energy, it is challenging to utilize and realize renewable energy due to the uncertain, intermittent and variable nature. In this article, we provide a taxonomy of the state-of-the-art research in applying renewable energy in cloud computing datacenters from five key aspects, including generation models and prediction methods of renewable energy, capacity planning of green datacenters, intra-datacenter workload scheduling and load balancing across geographically distributed datacenters. By exploring new research challenges involved in managing the use of renewable energy in datacenters, this article attempts to address why, when, where and how to leverage renewable energy in datacenters, also with a focus on future research avenues.

Turansitioning from fossil fuels to the use of renewable energy sources, such as solar and wind, to power cloud computing datacenters is a hot research topic today. In this article, we review and summarize a series of fundamental relevant research problems: why, when, where and how to leverage renewable (green or clean) energy in datacenters.

How Big is the Datacenter Power Consumption?

The proliferation of cloud computing services has promoted massive-scale, geographically distributed datacenters with millions of servers. Large cloud service providers consume many megawatts of power to operate such datacenters and the corresponding annual electricity bills are in the order of tens of millions of dollars — such as Google with over 1,120 GWh and $67 M, and Microsoft with over 600 GWh and $36 M [1]. Reportedly, datacenters now consume about 1.3 percent of the worldwide electricity and this fraction will grow to 8 percent by 2020 [2]. As a matter of fact, the aggregate datacenters in the world consumes more electricity than most nations in the world except 4 countries [3].

How Clean are the Current Datacenters?

High energy consumption not only results in large electricity cost, but also incurs high carbon emission. In the United States, generating 1 kWh of electricity emits about 500g of CO2 on average [2]. Each 100MW power station will cost $60-100 million dollars to build and emit 50 million tons of CO2 during its operation [3]. As a result, IT carbon footprints currently occupy 2 percent of global greenhouse gas emissions [4]. To measure how clean is a datacenter, the Green Grid organization proposes a new sustainability metric, carbon usage effectiveness (CUE), to measure carbon emissions associated with datacenters. CUE is defined as: $\text{CUE} = \frac{\text{Total CO2 Emissions caused by Total Datacenter Energy}}{\text{IT Energy Consumption}}$. The units of the CUE metric are kilograms of carbon dioxide equivalent (kgCO2e) per kilowatt-hour (kWh).

The comparison of carbon emission rate of the most common energy sources is shown in Table 1. We observe that the renewable energy sources have much less carbon emission rate than fossil fuels such as coal, gas and oil. Even though, fossil fuels still counts for 2/3 of electricity of the world [2].

Large IT companies have started to build datacenters with renewable energy, such as Facebook’s solar-powered datacenter in Oregon and Green House Data’s wind-powered datacenter in Wyoming. In April 2012, Greenpeace released a report asking, “How clean is your cloud?” [4] It examines the datacenters built by the large Internet companies, and ranks them according to how efficient their cloud facilities are, where they get their electricity. It found that both Google (39.4 percent clean energy) and Yahoo (56.4 percent clean energy) are active in supporting policies to drive renewable energy investment and to power clouds with green energy. In contrast, many large IT companies, such as Amazon, Apple and Microsoft, rapidly expand their cloud business without adequate attention on the electricity source, and rely heavily on brown energy to power their clouds. Even worse, there are numerous small and medium-sized datacenters that consume the majority of energy, yet much less energy efficient.
Why Leveraging Renewable Energy in Cloud Datacenters?

There have been numerous academic and commercial endeavors for reducing power costs and carbon footprints by applying better energy-proportional computing technologies, and more efficient power delivery and cooling systems. However, energy efficiency alone will slow down the growth of IT carbon footprint. To maintain safe levels of global greenhouse gases, renewable energy sources is becoming a prioritized choice for IT companies to power their rapidly expanding datacenter infrastructures. The international environmental organization Greenpeace states that “Green IT = Energy Efficiency + Renewable Energy” [4].

Thus, many governments enact renewable portfolio standards and provide incentives for green power generation and usage. For instance, California aims to generate 33 percent of its electricity from renewable sources by 2020 [5]. Federal and state incentives in New Jersey reduce the capital costs of renewable energy deployment by 60 percent [6]. Additionally, the improvements in power generating efficiency and cost/Watt reductions of renewable energy will reduce the deployment cost significantly in the future. For example, the efficiency of solar panels is expected to triple but the cost/Watt of solar panels is expected to halve in 2030 [6].

Apart from the pressures for reducing the huge energy consumption and carbon emission, there are unique opportunities raised by cloud datacenters to apply green energy:

- Cloud service providers typically own geographically distributed datacenters. They can distribute workloads among geo-dispersed datacenters to benefit from the location diversity of different types of available renewable energies.
- Cloud datacenters usually support a wide range of IT workloads, including both delay-sensitive non-flexible applications such as web browsing and delay-tolerant flexible applications such as scientific computational jobs. The workload flexibility can tackle the challenges in integrating intermittent renewable energy by delaying flexible workloads to periods when renewable sources are abundant without exceeding their execution deadlines.
- Cloud datacenters connect to grids at different locations with time-varying electricity prices. Renewable energy is a means to mitigate the risk against the future rise in power prices.
- Datacenters usually equip with uninterrupted power supply (UPS) in case of power outages. Since UPS battery is usually over-provisioned, UPS can store energy during periods of high renewable generation and supply power when the renewable energy is insufficient.

Pathway to a Cleaner Cloud

The most common way to utilize renewable energy is to deploy on-site generation equipments at the datacenter facility. Such on-site generation has negligible transmission and distribution loss. However, the location of a datacenter does not necessarily have a profitable on-site renewable energy deployment. Another way is to deploy off-site renewable energy generation plant at the locations with good wind speed or solar irradiation. The generated green energy can be dispatched across the grid to the consuming datacenters [3].

Apart from the above explicit options of provisioning renewable energy generation plants, there also exist three kinds of implicit options to utilize renewable energy:

- Power purchase agreement (PPA), which purchases a portion of green energy from a renewable energy source.
- Renewable energy credits (RECs), which are tradable, non-tangible energy commodities. They represent that 1 MWh of electricity be generated from an eligible renewable energy resource.
- Carbon offsetting, which represents the reduction of one ton of carbon dioxide. Table 2 shows the costs and carbon emission rate of REC, PPA conventional grid, and diesel generator (DG) used as backup power source in datacenters.

### Research Challenges

The main challenge of utilizing renewable energy is that renewable sources are variable, intermittent and unpredictable. For instance, wind or solar energy is only available when wind blows or sun shines. Characteristics of cloud datacenters further raise challenges to apply renewable energy:

- Global users require 7 × 24 cloud services: However, intermittent renewable energy presents a problem for consistent users of power like datacenters.
- Cloud demand is dynamic, which requires dynamic power provisioning. Thus, the power supply should be elastic. Nevertheless, unlike conventional energy from the grid, renewable energy can not be scheduled on demand.
- High reliability services. This incurs the problem of how to construct reliable power supply in the presence of uncertain renewable energy.
- Automatic management. This requires the power supply system should choose and supply power automatically among multiple power sources.

Before deploying renewable energy equipment, we need to know what kinds of datacenter workloads are amenable for green energy and whether the datacenter is suitable for green energy. Afterwards, we need to know the capacity of renewable energy generation for the specific datacenter demand. To mitigate the variability of renewable energy, datacenters can store the generated renewable energy in batteries or on the grid. However, it remains significant challenges when/how to store and supply energy from the battery to meet demand. In addition, these approaches incur energy loss and high additional costs of purchase and maintenance. Instead, we can maximize the use of available energy by matching the demand to the supply. This prompts many interesting research problems. Due to the variable availability of renewable energy, the scheduler has to exploit suitable forecast techniques to predict the available renewable energy in an online fashion. How can we schedule workload in intra-datacenter to maximize renewable energy usage? Is it profitable to leverage geographical workload for energy minimization? We will answer these questions in next sections.

### Table 1. Carbon emission rate of major energy sources [2, 3].

<table>
<thead>
<tr>
<th>Energy Source</th>
<th>Nuclear</th>
<th>Coal</th>
<th>Gas</th>
<th>Oil</th>
<th>Hydro</th>
<th>Wind</th>
<th>Solar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon Emission Rate (gCO₂e/kWh)</td>
<td>15</td>
<td>968</td>
<td>440</td>
<td>890</td>
<td>13.5</td>
<td>22.5</td>
<td>53</td>
</tr>
</tbody>
</table>

### Table 2. Comparison of costs and prices of energy sources [3].

<table>
<thead>
<tr>
<th>Energy Source</th>
<th>Grid</th>
<th>PPA</th>
<th>REC</th>
<th>DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost (¢/kWh)</td>
<td>5</td>
<td>6</td>
<td>0.5</td>
<td>30</td>
</tr>
<tr>
<td>Carbon Emission Rate (gCO₂e/kWh)</td>
<td>586</td>
<td>0</td>
<td>0</td>
<td>1056</td>
</tr>
</tbody>
</table>
Energy Real-time, Solar, Web request Energy-proportional Prediction, Technique Data Wind, grid Web request iSwitch 

### Table 3. A taxonomy of approaches of leveraging renewable energy in datacenters.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Energy Source</th>
<th>Time</th>
<th>Workload</th>
<th>Level</th>
<th>Input</th>
<th>Evaluation Method</th>
<th>Goa</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity Planning [3]</td>
<td>On/off-site, ESD/grid</td>
<td>Off</td>
<td>Pervasive</td>
<td>Data center</td>
<td>Historical Prediction</td>
<td>Model, trace-driven simulation</td>
<td>Optimal energy portfolio to minimize cost</td>
<td>Modeling different renewables, carbon target, policy, tax, ESD</td>
</tr>
<tr>
<td>ReRack [12]</td>
<td>Wind, solar, ESD/grid</td>
<td>Off</td>
<td>Interactive</td>
<td>Data center</td>
<td>Historical Prediction</td>
<td>Model, ReRack simulator</td>
<td>Evaluate different energy portfolio cost</td>
<td>Modeling different renewables, ESD, weather, incentive</td>
</tr>
<tr>
<td>iSwitch [16]</td>
<td>Wind, grid</td>
<td>On</td>
<td>Web request</td>
<td>Data center</td>
<td>Real-time</td>
<td>iSwitch prototype</td>
<td>Maximize renewables usage, performance</td>
<td>VM migration</td>
</tr>
<tr>
<td>Prediction [10]</td>
<td>Solar, wind, grid</td>
<td>On</td>
<td>Web request batch</td>
<td>Data center</td>
<td>Prediction</td>
<td>Model, simulation</td>
<td>Adapt server load with renewables</td>
<td>Two job queues, Defer batch tasks</td>
</tr>
<tr>
<td>Greening [18, 19]</td>
<td>Solar, wind</td>
<td>On</td>
<td>Interactive</td>
<td>Inter-DC</td>
<td>Prediction, real-time</td>
<td>Model, trace-driven</td>
<td>Using entirely renewable energy</td>
<td>Follow the renewables routing</td>
</tr>
<tr>
<td>Not Easy Being Green [2]</td>
<td>Fossil fuel, wind, hydro</td>
<td>On</td>
<td>Web request</td>
<td>Inter-DC</td>
<td>Real-time</td>
<td>FORTE, trace-driven</td>
<td>Tradeoff delay, energy cost and carbon footprint</td>
<td>Renewable energy aware requests routing</td>
</tr>
<tr>
<td>GreenWare [8]</td>
<td>Solar, wind, grid</td>
<td>On</td>
<td>Web request</td>
<td>Inter-DC</td>
<td>Real-time, prediction</td>
<td>Simulator, trace-driven</td>
<td>Maximize usage of available renewables under budget</td>
<td>Budget, fractional linear programming</td>
</tr>
<tr>
<td>Free Lunch [17]</td>
<td>Solar, wind</td>
<td>On</td>
<td>All VM application</td>
<td>Inter-DC</td>
<td>Real-time</td>
<td>Free Lunch framework</td>
<td>Fully use available renewable energy</td>
<td>VM migration</td>
</tr>
<tr>
<td>Some Joules [13]</td>
<td>Solar, wind, grid</td>
<td>On</td>
<td>Web request</td>
<td>Inter-DC</td>
<td>Real-time, prediction</td>
<td>Model, simulation</td>
<td>Fully use available renewable energy</td>
<td>Renewable energy aware requests routing</td>
</tr>
</tbody>
</table>

### Existing Approaches

Recently, various advanced solutions have been proposed to address the above challenges, which can be broadly classified into five topics: generation model and prediction model of renewable energy, capacity planning of green datacenter, intra-datacenter workload scheduling and inter-datacenters load balancing. Table 3 summarizes the taxonomy of existing approaches.

#### Renewable Energy Generation Model

Wind and solar are the most prominent renewable sources, which currently provide 62 percent and 13 percent non-hydro renewable electricity worldwide, respectively [3]. The solar or wind power output depends on the environmental conditions, such as solar irradiance or wind speed. Consequently, their capacity factor, which is the ratio of the actual output over a period to its potential output if it operated at full nameplate capacity, can be much lower than grid energy (e.g., 80 percent). Specifically, the capacity factor of wind energy is within 20–45 percent, while the capacity factor of solar energy ranges from 14–24 percent [3].

The wind turbines generation can be modeled as a function of the wind speed. Let $v$ be the wind speed, then the wind power output $P_{wind}$ can be approximated as follows [8]:

$$P_{wind} = \begin{cases} 
0 & v < v_{in}, v > v_{out} \\
\frac{v - v_{in}}{v_{r} - v_{in}} & v_{in} < v < v_{r} \\
\frac{v_{r} - v_{in}}{v_{r} - v_{out}} & v_{r} < v < v_{out} 
\end{cases}$$

where $v_{in}$ and $v_{out}$ are the cut-in (typically 3–4 m/s) and cut-out (typically 25 m/s) wind speed, and $v_{r}$ and $p_{r}$ are the nameplate speed and wind turbines power. If a wind farm consists of $m_{turb}$ turbine, then the wind power output is the sum of all the turbine power output:

$$PW = \sum_{k=1}^{m_{turb}} P_{wind}$$

where $P_{wind}$ is the power output of $k$th wind turbine at the wind speed of $v$, with the assumption that the wind turbines have the same wind speed in the same wind farm. However, the speed and angle may differ from turbine to turbine in a farm. A more practical model should be, where $v_k$ is the wind speed that is a vector containing the wind direction.

The solar photovoltaic (PV) generation can be modeled as its current-voltage characteristic equations, which are related to solar cell parameters and environmental conditions, such as irradiance condition $G$ and temperature $T$ [8]:

50
Figure 1. Overall vision.

\[
i(G,T) = I_{dc}(G,T) - I_{oc}(T) \cdot e^{-\frac{V_{os} k_i (T - T_0)}} \\
\]

where short-circuit current

\[
I_{sc}(G,T) = \frac{G}{G_0} I_{dc}(1 + \frac{k_1}{100} (T - T_0)) \\
\]

dark saturation current

\[
I_{ds}(T) = I_{ds}(1 + \frac{k_2}{100} (T - T_0)) \cdot e^{-\frac{V_{os} k_i (T - T_0)}} \cdot V_{dh} / q \\
\]

is the junction thermal voltage and \( k \) is Boltzmann’s constant. \( T \) is the ambient temperature and \( q \) is the charge of the electron. \( G_0 \) and \( T_0 \) are the irradiance level and temperature in Standard Test Conditions (STC), i.e., \( G_0 = 1000\text{W/m}^2, T_0 = 25^\circ\text{C} \). \( I_{dc}, V_{dc}, k_1 \) and \( k_2 \) are the rated parameters of short-circuit current, open-circuit voltage, temperature coefficients of the open-circuit and the short-circuit in STC, respectively.

The solar power output can be estimated as the maximal power point (mpp) that can be extracted from the PV panel \[9\], i.e., \( \text{mpp} = I_{mp} \cdot r_{mp} \) where \( r_{mp} \) is the optimal load. \( I_{mp} \) is the corresponding current,

\[
r_{mp} = R_s + \frac{n_s \cdot V_{sh}}{I_{dc}(G,T) + I_{oc}(T) - I_{mp}} \\
\]

and \( n_s \) is the number of the solar cells in the PV panel connected in series, e.g., \( n_s = 72 \). Thus, if there are \( n_s \) PV panels in the solar plant, the overall solar power output is estimated as: \( PS = \sum_{k=1}^{n_s} \text{mpp}^k \), where \( \text{mpp}^k \) is the maximal power value from the \( k \)th PV panel with irradiance condition \( G \) and temperature \( T \).

**Renewable Energy Prediction Model**

Since renewable energy is highly variable, it is important to predict the available energy when scheduling workloads to match renewable energy supply. Solar energy prediction typically applies estimated weighted moving average (EWMA) models, for its relative consistency and periodic patterns. However, with frequent weather changes, the prediction mean error is over 20 percent \[10\]. Another prediction model is weather-conditioned moving average (WCMA), which takes into account the mean value across days and a solar condition factor in the present day relative to previous days \[10\]. Differently, load-scheduling system GreenSlot \[7\] and GreenHadoop \[11\] predict the energy generation as a function inversely related to the amount of cloud coverage. That is, \( E_p(t) = B(t)(1 - \text{CloudCover}) \), where \( E_p(t) \) is the estimated solar energy for time \( t \), \( B(t) \) is the ideal solar energy generation, and CloudCover is the estimated percentage of cloud coverage compared to sunny days (within \( 0~1 \)).

Wind energy forecast typically has two major approaches: time-series analysis of power data and wind speed prediction. Since the response time constraints of cloud services can be quite short, the predictor needs to be fast. Aksanli et al. \[10\] developed a low-overhead wind energy predictor that utilizes available data to generate a power wind table based on wind velocity and wind direction as follows:

\[
P_{\text{new}}(v, d) = \alpha \times P_{\text{obs}}(v, d, t) + (1 - \alpha) \times P_{\text{old}}(v, d) \\
\]

where \( P_{\text{new}}(v, d) \) is the new power value with a wind velocity \( v \) and direction \( d \), \( P_{\text{old}}(v, d) \) is the existing value for the same velocity and direction, and \( P_{\text{obs}}(v, d, t) \) is the measured value at time \( t \). \( \alpha \) is within \( 0~1 \). When predicting future interval wind energy, it does a table lookup with the predicted wind velocity and direction: \( P_{\text{pred}}(v, d, t + k) = P(v(t + k), d(t + k)) \). For a 30-minute prediction interval, the predictor shows a mean error of 17.2 percent, no worse than the time-series models \[10\].

**Capacity Planning of Green Datacenter**

Renewable energy can reduce datacenter operational cost and carbon footprint when correctly selected. Nevertheless, as illustrated in Fig. 1, it is not a trivial task, for the reason that datacenter planners should consider different renewable options (on-site, off-site, PPA, REC), different energy sources (solar, wind), different energy storage devices (batteries, fuel-cells, flywheels), markets prices fluctuations, workload variations, weather conditions, incentives and service agreement penalties. Given all these factors, Jose Renau et.al. \[12\] proposed ReRack, a power simulator that can evaluate the energy cost of a datacenter using renewable energy. For any given location and workload of a datacenter, ReRack can find the best ratio of renewable energy sources using a genetic algorithm.

Differently, Chuangang Ren et al. \[3\] developed a practical linear programming based optimization framework for a datacenter to achieve a target carbon footprint at minimal cost, among the various options presented in Fig. 1. Under extensive experiments with real-world power demand and renewable energy traces, the key findings are:

- Renewable energy can lower both carbon emissions and costs for datacenters.
- On-site renewables can lower datacenter costs by reducing the peak power draw from grid.
- The most cost-effective approaches for carbon reduction vary with different carbon footprint targets.

For a moderate target (up to 30 percent), the best options is using on-site renewables, while a more carbon reduction goal requires off-site renewables, and a zero carbon goal must resort to renewable energy products such as RECs. Unfortunately, the model considered a low and constant grid price ($0.05/kWh), which made on-site generation relatively expensive in comparison.
— To study the use of renewable energy in job scheduling power state management and load migration. In a modern datacenter with on-site renewable energy, there are three approaches to match power demand to variable renewable energy, with respect to job scheduling, server power state management and load migration.

**Intra-Datacenter Workload Scheduling**

In a modern datacenter with on-site renewable energy, there are three approaches to match power demand to variable renewable energy, with respect to job scheduling, server power state management and load migration.

**Job Scheduling** — To study the use of renewable energy in datacenters, a renewable energy powered platform “Parasol” was constructed [6]. Two load-scheduling systems GreenSlot [7] and GreenHadoop [11] were designed. Both assume that the datacenter is powered by solar energy and grid with no batteries for energy storage. GreenSlot [7] is for batch jobs while GreenHadoop [11] is for Hadoop jobs. The goal is to maximize the use of solar energy while meeting the jobs’ deadline. They predict the available solar energy in the future using historical data and weather forecasts, and then defer workload execution until solar energy is available without violating the deadline. If the grid energy must be used to avoid deadline violations, the systems schedule jobs for times when grid energy is cheap. To demonstrate these in practice, various experiments are conducted in the micro-datacenter Parasol [6] with 16-node powered by a solar array and the electrical grid (as a backup). GreenSlot is robust to different workload characteristics, providing green energy increases of at least 19 percent and energy cost savings of at least 25 percent compared to backfilling scheduling algorithm. The solar energy predictor is reasonably accurate, achieving hour-ahead median and 90th percentile errors of 12.9 percent and 24.6 percent, respectively. Interestingly, they found that even perfect knowledge of solar energy availability increases solar energy use and decreases cost both by only 1 percent.

GreenSlot and GreenHadoop are not suitable for interactive and delay-sensitive workloads. Aksanli et al. [10] designed an adaptive scheduler for mixed batch and Web service jobs. The scheduler uses two separate job arrival queues: one for the Web services and the other for the batch jobs. Each server poses a limit on the amount of workloads (Web requests and batch jobs) to be served. Web services start execution whenever there are available computing resources, while batch jobs start execution based on the availability of the renewable energy. Similarly, an energy agile computing cluster was proposed [14], in which it also slightly defers long running batch jobs so as to make the best use of wind energy. In addition, it enables graceful degradation of interactive services to adapt power consumption.

**Server Power State Magement** — SolarCore [9] is a solar energy driven, multi-core architecture power management scheme that moves the solar array operating point close to the maximum power point under changing atmospheric conditions. SolarCore performs load adaptation utilizing dynamic voltage/frequency scaling (DVFS) and per-core power gating (PCPG). It intelligently allocates the dynamically varying power budget across multiple cores to maximize workload performance. When the solar power supply drops below a certain threshold, it switches to grid. Evaluations with solar traces and SPEC2000 benchmark show that SolarCore boosts multi-core processor performance by 43 percent compared with fixed power budget scheme and 10.8 percent compared with round-robin load adaptation. Researchers from UMass at Amherst have built Blink [15], a cluster of 10 laptop motherboards powered by two micro wind turbines and two solar panels. They used batteries only as a small 5-minute energy buffer. Blink modulates the server power state using fast sleep to adapt to the availability of renewable energy. Yet, the proposed cluster is not connected to the grid, which seems to be unrealistic. Datacenters that completely depend on variable green energy can cause unbounded performance degradation.

**Load Migration** — With the presence of the intermittent renewable power, inefficient and redundant workload matching activities can incur up to 4X performance loss [16]. iSwitch [16] explores the design tradeoffs between load tuning overhead and green energy utilization in renewable energy driven datacenters. Servers are divided into two groups: one is powered by grid while the other is powered by on-site wind energy. Based on the availability of wind energy, iSwitch migrates load between these two groups to best utilize the renewable energy generation. iSwitch can mitigate average network traffic by 75 percent, peak network traffic by 95 percent, and reduce 80 percent job waiting time while still maintaining 96 percent renewable energy utilization.

**Inter-Datacenters Load Balancing**

Large cloud service providers usually own geo-distributed datacenters for:

- Better performance, i.e., close to end users
- More reliable system, i.e., multiple backups in different datacenters
- Lower cost, i.e., leveraging different energy sources, prices, cooling efficiency and tax

The first question is whether geographical load balancing can exploit renewable energy and reduce environmental impact? FreeLunch was proposed to exploit renewable energy [17]. It assumes that datacenters are close to renewable energy sources, and support the seamless execution and migration of virtual machines according to the power availability. High latency and storage synchronisation can degrade the efficiency of VM migration. Thus, FreeLunch might be inefficient for latency sensitive or interactive applications.

For web requests, Liu et al. exploited whether geographical load balancing can encourage the use of renewable energy and reduce the use of fossil fuel (brown) energy [18]. It was found that if grid energy (with penetration of renewable energy sources) is dynamically priced based on the proportion of

---

**Figure 2. The platform provider GUI.**
brown energy, then geographical load balancing can reduce brown energy usage significantly. Especially, the optimal pricing model at ith datacenter is: \( p_i(t) = (1 - \alpha_i(t))/\beta \), where \( \alpha_i(t) \) the fraction of the energy that is from renewable sources at time \( t \) and \( \beta \) is the relative valuation of delay versus energy. Further, based on trace-driven study, they investigated the feasibility of powering datacenters entirely using renewable energy [19]. It observed that “follow the renewables” routing could significantly reduce the required capacity of renewable energy, especially in corporation with suitable size of energy storage. Storage becomes more valuable with higher capacities of renewables, i.e., >1.5, where renewable capacity is defined as the ratio of energy required to serve the average workload. Moreover, they found that wind energy is more viable than solar. The rational is that wind is available during both day and night, and has little correlation across different locations.

The second question is how many and which user requests should be migrated? Stewart et al. [13] proposed an accounting model of per-request energy consumption as a guideline for managing renewable energy. Based on the performance counters, the model collects three metrics including the activities and power model is given as:

\[
p_i(t) = \left(1 - \alpha_i(t)/\beta\right)
\]

where \( p_i^* \) are the coefficient parameters for the linear model, and \( C_i^{\text{ceil}} \) are the upper bound for the three metrics. With the model, the power/energy profiles for individual requests can be constructed, and renewable-aware request distribution can be conducted.

If incentives for renewable energy usage are not considered, currently renewable energy is often more expensive than brown energy. To utilize renewable energy, cloud service providers have two challenges:

- How to conduct dynamic request dispatching among geographical datacenters to maximize the use of renewable energy?
- How to achieve that within allowed operational cost budget.

As illustrated in Fig. 2, GreenWare [8] dynamically distributes requests among datacenters to maximize the percentage of used renewable energy subject to the desired cost budget. This is formulated as a constrained optimization problem:

\[
\text{Maximize:} \quad \sum_{i=1}^{N} W_i + S_i
\]

\[\text{s.t.} \quad \sum_{i=1}^{N} \left( P_{W_i} W_i + P_{S_i} S_i + P_{B_i} B_i \right) \leq \text{Budget}\]

where \( W_i, S_i, B_i \) are the amount of used wind, solar and brown energy in the ith (1 \( \leq i \leq N \)) datacenter, and \( P_{W_i}, P_{S_i}, \) and \( P_{B_i} \) are the current prices of the three types of energy. Constraint (6) means that the energy cost should be within the cost budget. They proposed an efficient request-dispatching algorithm based on linear-fractional programming (LFP) and implemented GreenWare based on the linprog solver in Matlab. Evaluations with real-world weather, electricity price and workload traces show that GreenWare can significantly increase the use of renewable energy without violating the desired cost budget.

However, GreenWare ignores a key factor: what set of data users are interested in and where the data is? FORTE [2] captures the relationship between user groups, datacenters and data. It uses two assignment algorithms to optimally map users to datacenters. Basers and map data to datacenters. The objective is to minimize the weighted sum of access latency, electricity cost and carbon footprint. Trace-driven experiments show that FORTE can reduce carbon emissions by 10 percent without increasing the electricity bill and the mean latency.

### Moving Forward: Future Research Trends

#### Diverse Green Sources

Most existing works focus on studying the usage of solar or wind energy. However, many other renewable energy sources are ignored, such as tidal energy, hydrogen energy and fuel cells. Especially, fuel cells can generate energy with steady power supply, which is very different from the intermittent solar or wind energy, and thus require different research methods. Besides, off-site renewables such as PPA and REC shown in Fig. 1, also deserve deeper theoretical study and practical application, so as to find the best energy portfolio for a specific datacenter.

#### New Power Supply Architectures

If constructing on-site renewable energy sources, how to integrate the renewable energy into the traditional grid circuit? There are two common ways:

- **Transfer switches.** Modern datacenters can automatically switch power sources via Automatic Transfer Switch (ATS) based on a pre-configured power-transfer threshold. If the power supply from the primary source falls below a threshold, the ATS switches to a secondary energy source.
- **Grid ties:** Grid ties combine electricity produced from the renewable sources and the grid on the same circuit for power supply. However, transfer switches isolate electricity produced on site, keeping it separated from the electricity grid. In contrast, grid ties allow multiple sources to power a server. If a new datacenter power supply system desires to isolate variable renewable energy from the stable electricity grid, the transfer switch is a better choice. The power-transfer threshold computation and grid-tie placement are therefore also important research problems. Further, unlike current centralized cloud model, we can consider peer-to-peer-based architecture to leverage the distributed renewable energies.
Demand Side Workload Management

Current works either solely consider delay-sensitive demand or only suit for delay-tolerant workloads, while others neglect the interaction among uncertain renewable energy generation, energy storage management and time-varying electricity prices. It remains a significant challenge how to operate among multiple power supply sources in a complementary manner to deliver reliable energy to datacenter users with arbitrary demand over time. As shown in Fig. 3, one possible method is to address delay-sensitive demand when it is generated, and cater to the delay-tolerant demand in a more strategic fashion over time, while guaranteeing that it is served before the deadline. Energy storage can store renewable energy and cheap grid electricity, while supply power when needed, so as to best utilize the available renewable energy and the periods with lower electricity prices from the grid. Another way to address the demand and supply uncertainties is demand response mechanism, which is a strategy for achieving energy efficiency through managing power consumption in response to supply conditions. For example, dynamic electricity pricing and peak-demand charging can induce end users to curb their demand at critical times in an effort to reduce the costs. However, it is still challenging to design adequate pricing mechanisms to reflect the availability of renewable energy.

Supply Side Power Management

The availability of renewable energy is time-varying. Energy storage devices can alleviate this variability. However, energy storage devices incur energy losses due to self-discharging and internal resistance. The related costs of energy storage devices may dominate in systems powered by renewable energy. It is thus a non-trivial problem. With a plethora of energy storage technologies available today, we need to model the tradeoffs in cost, density, lifetime and efficiency when provisioning energy storage. In addition, most current datacenters use centralized UPS batteries. Newer datacenters are starting to use decentralized batteries, e.g., rack-level in Facebook or server-level in Google. The centralized vs. distributed and hierarchical placement problems of energy storage devices deserve thorough investigation. Moreover, today most datacenters have an inefficient power conversion. Power is delivered using alternating current (AC) that goes through multiple conversions between AC and direct current (DC), resulting in energy loss up to 30 percent. Future DC-based power distribution systems can improve renewable energy utilization and reduce the total system energy use.

Conclusion

The main challenge of utilizing renewable energies is the variable, intermittent and unpredictable nature. By presenting a taxonomy of the latest research and exploring new challenges involved in managing the use of uncertain renewable energy, we intend to answer why, when, where and how to leverage renewable energy in datacenters. Specifically, we believe that matching uncertain power demand and multi-sources supply in a complementary manner should be one of the highlights in future research.

Acknowledgment

The corresponding author is Fangming Liu (fmliu@hust.edu.cn). The research was supported in part by a grant from National Basic Research Program (973 program) under Grant of 2014CB347800, by a grant from The National Natural Science Foundation of China (NSFC) under grant No.61133006 and No.61103176, by a grant from the Research Fund of Young Scholars for the Doctoral Program of Higher Education, Ministry of Education, China, under grant No.20110142120079, by the CHUTIAN Scholar Project of Hubei Province. Prof. Bo Li's work was supported in part by a grant from RGC under the contract 615613, a grant from NSFC/RGC under the contract N_HKUST610/11, and a grant from ChinaCache Int. Corp. under the contract CCNT12EG01.

References


Biographies

WEI DENG (wdeng@hust.edu.cn) is currently a Ph.D. student in the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. He was the recipient of Best Paper Nominee from IEEE CloudCom 2012. His research interests focus on cloud computing, datacenter networking, green computing, smart grids, modeling and optimization.

FANGMING LIU (M) (fmliu@hust.edu.cn) is an associate professor in the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China; he is awarded the CHUTIAN Scholar of Hubei Province, China. He is the Youth Scientist of National 973 Basic Research Program Project on Software-defined Networking (SDN)-based Datacenter Networks, which is one of the largest SDN projects in China. Since 2012, he has also been invited as a StarTrack Visiting Young Faculty in Microsoft Research Asia (MSRA), Beijing. He received his B.Eng. degree in 2005 from Department of Computer Science and Technology, Tsinghua University, Beijing; and his Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology in 2011. From 2009 to 2010, he was a visiting scholar at the Department of Electrical and Computer Engineering, University of Toronto, Canada. He was the recipient of...
two Best Paper Awards from IEEE GLOBECOM 2011 and IEEE CloudCom 2012, respectively. His research interests include cloud computing and data-center networking, mobile cloud, green computing and communications, software-defined networking and virtualization technology, large-scale Internet content distribution and video streaming systems. He is a member of IEEE and ACM, as well as a member of the China Computer Federation (CCF) Internet Technical Committee. He has been a Guest Editor for IEEE Network Magazine, an Associate Editor for Frontiers of Computer Science, and served as TPC for IEEE INFOCOM 2013-2014 and GLOBECOM 2012-2013.

HAI JIN [SM] (hjin@hust.edu.cn) is a Cheung Kung Scholars Chair Professor of computer science at the Huazhong University of Science and Technology (HUST), China. He is now dean of the School of Computer Science and Technology at HUST. He received his Ph.D. degree in computer engineering from HUST in 1994. In 1996, he was awarded a German Academic Exchange Service fellowship to visit the Technical University of Chemnitz in Germany. He was awarded the Excellent Youth Award from the National Science Foundation of China in 2001. He is the chief scientist of ChinaGrid, the largest grid computing project in China, and chief scientist of the National 973 Basic Research Program Project of Virtualization Technology of Computing Systems. His research interests include computer architecture, virtualization technology, cloud computing and grid computing, peer-to-peer computing, network storage, and network security.

BO LI [F] (bli@cse.ust.hk) is a professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His current research interests include large-scale content distribution in the Internet, datacenter networking, cloud computing, and mobile sensor networks. He made pioneering contributions in the Internet video broadcast with a system, called Coolstreaming (the keyword had over 2,000,000 entries on Google), which was credited as the world first large-scale Peer-to-Peer live video streaming system, which spearheaded a momentum in video broadcast industry, with no fewer than a dozen successful companies adopting the same mesh-based pull streaming technique to deliver live media content to hundreds of millions of users in the world. He has been an editor or a guest editor for a dozen of IEEE journals and magazines. He was the Co-TPC Chair for IEEE INFOCOM 2004. He received his B. Eng. Degree in the Computer Science from Tsinghua University, Beijing, and his Ph.D. degree in the Electrical and Computer Engineering from University of Massachusetts at Amherst.

DAN LI [M] (tolidan@tsinghua.edu.cn) is currently an associate professor in the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include Internet architecture and protocols, cloud computing and data center networking.