Thermal-Aware Hybrid Workload Management in a Green Datacenter towards Renewable Energy Utilization

Yuling Li [1], Xiaoying Wang [1],*, Peicong Luo [1] and Qingyi Pan [1,2]

[1]   State Key Laboratory of Plateau Ecology and Agriculture, Department of Computer Technology and Applications, Qinghai University, Xining 810016, China; lyl_feng@qhu.edu.cn (Y.L.); peicongluo@qhu.edu.cn (P.L.); panqingyi@qhu.edu.cn (Q.P.)
[2]   Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
*   Correspondence: wxy_cta@qhu.edu.cn; Tel.: +86-0971-5311571

check for
updates

Abstract: The increase in massive data processing and computing in datacenters in recent years has resulted in the problem of severe energy consumption, which also leads to a significant carbon footprint and a negative impact on the environment. A growing number of IT companies with operating datacenters are adopting renewable energy as part of their energy supply to offset the consumption of brown energy. In this paper, we focused on a green datacenter using hybrid energy supply, leveraged the time flexibility of workloads in the datacenter, and proposed a thermal-aware workload management method to maximize the utilization of renewable energy sources, considering the power consumption of both computing devices and cooling devices at the same time. The critical knob of our approach was workload shifting, which scheduled more delay-tolerant workloads and allocated resources in the datacenter according to the availability of renewable energy supply and the variation of cooling temperature. In order to evaluate the performance of the proposed method, we conducted simulation experiments using the *Cloudsim-plus* tool. The results demonstrated that the proposed method could effectively reduce the consumption of brown energy while maximizing the utilization of green energy.

Keywords: green datacenter; renewable energy; workload management; power consumption; thermal-aware

1. Introduction

In recent years, with the rapid spreading and development of cloud computing technology around the world, a large number of computing operations in the datacenters need to respond rapidly and efficiently to ensure service capabilities. However, the growing demand for cloud infrastructure has led to a dramatic increase in the power consumption of datacenters, which has become a significant issue need to be solved. Datacenters around the world consume a lot of energy each year, and the average power consumed by each datacenter is almost equal to the power consumed by 25,000 homes in the United States [1]. In 2017, there were approximately 8 million datacenters around the world that consumed 416.2 terawatt hours of electricity [2]. This is equivalent to 2% of the total electricity consumption in the world and is expected to reach 5% of global electricity consumption by 2020. As estimated, the energy cost of a datacenter approximately accounts for nearly 50% of the total operating cost of the datacenter. This results in most of the total electrical energy consumed not being sufficiently utilized. One of the major reasons for this is that the datacenter has a certain proportion of idle energy consumption during its operation. Even at very low utilization rates, such as 10% CPU usage, the power consumed exceeds 50% of the peak power [3]. Some methods that dynamically

migrate or consolidate tasks on some less-utilized servers or turn off idle servers have been proven to be energy-efficient strategies [4–6]. However, most of these studies only considered the computing power consumption of the datacenter or the cooling power consumption separately, without combining them together.

In recent years, the use of green energy sources such as solar energy, wind energy, and tidal energy has become a global trend in building green sustainable datacenters [7–9]. Kong et al. [10] conducted a survey on the renewable energy used and carbon emission in many datacenters. Compared with brown energy, green energy has lots of advantages, such as being natural, renewable, clean, and low cost. However, the generation of the renewable energy such as wind, solar, and tidal energy are usually intermittent and unstable. Hence, a way to accurately predict the amount of available renewable energy is worth being studied in order to make full use of renewable sources in the datacenters. In recent years, famous IT companies such as *Microsoft*, *Google*, *Amazon*, and *IBM* were all operating large-scale datacenters around the world to cope with the growing computation demand, and they were also trying to use the renewable energy as partial supply to their datacenter to further reduce the energy cost. Therefore, a way to effectively manage such energy supply in the datacenter becomes an important issue for these service providers.

Nowadays, most datacenters support various types of workloads, including critical interactive workloads and batch-type workloads, wherein the latter can be deferred for a certain time to be processed. In general, interactive workloads include web browsing, real-time gaming, data query, and other workloads which need an immediate response. In contrast, batch-type workloads like image processing, scientific applications, and financial data analysis can be scheduled later as long as they could be completed before their deadlines [11]. This provides the feasibility of scheduling workloads in the datacenter in the time dimension.

The main objective of this paper is to manage workloads effectively in a green datacenter, aiming at making full use of renewable energy and minimizing the total power cost of the datacenter. In this paper we adopt solar energy as the renewable energy supply in the experiments. Due to fluctuant energy input and dynamically changing workload over time, we adjusted the number of workloads in each time slot and the temperature supplied by the cooling device to maximize the use of renewable energy. In this way, the brown power consumption of both IT devices and cooling devices could be decreased. Moreover, we adopted the neural network model to predict the amount of solar power generated to facilitate more accurate workload allocation decisions. This paper is an extended version of our prior work [12]. The biggest difference between this paper and the previous version is that we considered two types of workloads at the same time (interactive workloads and batch-type workloads). In addition, due to the unstable solar power generation, we conducted a prediction of the amount of solar energy generated in advance to better schedule the batch-type jobs, set more methods for comparison, and produce a more detailed analysis of results. And the number of words has increased by 50%.

The remainder of this paper is organized as follows. Section 2 introduces some related work on datacenter energy management. Section 3 presents the problem definition and the model used in this paper. Section 4 depicts the architecture of the green datacenter and the solar power prediction method. Section 5 describes the methods and strategies we designed to solve the defined problem. Section 6 analyzes the experiment results by comparing three different strategies. At last, Section 7 concludes the whole paper and discusses possible future work.

2. Related Work

In recent years, the establishment of green datacenters and the use of renewable energy have become a hot research topic. The main ideas of these studies are to save energy in datacenters from different aspects. For example, research has been carried out on resource management, scheduling of virtual machines, and balancing load management in datacenters.

2.1. Energy Saving Approaches in Datacenters by Performance Adjustment

In response to the high energy consumption in datacenters, some researchers have proposed various methods to reduce the energy consumption of datacenters from different aspects. Liu et al. [13] analyzed the composition of energy consumption of cloud computing datacenters and analyzed related research on energy management in cloud datacenters. They put forward that the energy consumption of the datacenter is composed of about 50% of the computing energy consumption and 40% of the air conditioning refrigeration energy consumption. The remaining 15% of energy consumption is mainly consumed by storage equipment and power distribution systems. This means that about half of the energy consumption of the datacenter is consumed by non-computing devices. Zhang et al. [14] studied the minimization problem of datacenter energy consumption that satisfies quality of service (QoS) and server CPU average temperature constraints. They proposed an energy minimization algorithm based on Lyapunov optimization theory to reduce the total energy consumption of the datacenter. Salma et al. [15] presented a non-traditional optimization technique, which minimizes the execution time and in turn reduces computation cost. Eduardo et al. [16] proposed methods to save energy by dynamically controlling the switches of servers in the cluster, and also proposed a series of methods for balanced or unbalanced loads. Sungkap et al. [17] proposed an ambient temperature-aware capping to maximize power efficiency while minimizing overheating. Wang et al. [18] put forward an analytical model, which describes datacenter resources with heat transfer properties and workloads with thermal features. Then, they proposed two thermal-aware tasks-scheduling algorithms that aim to lower temperatures and cooling system power consumption in a datacenter. Luo et al. [19] proposed an IT workload management method to process delay-tolerant jobs that have the same deadline of completion while maintaining their quality of service (QoS). Their approach has a workload shaping stage, which decides to admit the workload, and a scheduling stage, which aims to minimize the electricity costs of the datacenter; however, they did not include the power of the cooling system. Overall, methods to reduce energy consumption mainly include the migration and consolidation of virtual machines, powering off the host when the host is idle, and the delayed execution of tasks. Moreover, their results prove that the proposed methods can reduce the energy consumption of the datacenter to a certain extent.

2.2. Energy Saving Approaches in Datacenters by Virtual Machine Consolidation

Consolidating or migrating virtual machines can also save energy in the datacenter. In order to reduce energy consumption and ensure high compliance with service level agreements, Fahimeh et al. [4] proposed a reinforcement-learning-based dynamic consolidation method to minimize the number of active hosts according to the current resource requirement. Beloglazov et al. [20] proposed an adaptive heuristic algorithm for VM (Virtual Machine) consolidation based on the analysis of the historical data of VM resource utilization. Atefeh et al. [21] proposed two online deterministic algorithms, migrating virtual machines to the nearby datacenters with surplus renewable energy in order to save brown energy. Islam et al. [22] proposed a novel resource management algorithm, to optimally control the server capacity provisioning, virtual machine CPU allocation, and load distribution for minimizing the datacenter power consumption while satisfying the quality of service, IT peak power, and maximum server temperature constraints. Previous work [23] presented a novel VM scheduling mechanism for reducing the energy consumption of datacenter. They considered both load-balance and thermal-awareness to achieve the goal. Qouneh et al. [24] proposed to borrow virtual computing resources from GPU VMs and reallocate them to CPU VMs. In this way, they can minimize server power cost while maintaining overall server performance. This method is also a way to adjust the load of the datacenter to achieve energy saving goals. Prior work [25] defined and developed a set of performance and energy-aware strategies for resource allocation, task scheduling, and for the hibernation of virtual machines. They combined energy and performance-aware scheduling policies in order to transfer virtual machines into an idle state and the efficiency achieved by applying the proposed models has been tested using a realistic large-scale cloud computing system simulator.

Therefore, the consolidation and migrating of virtual machines also has an impact on the energy consumption of the datacenter.

2.3. Approaches to Improve the Energy Efficiency

There are also some researchers who improve energy efficiency by adjusting datacenter energy usage strategies. Dami'an et al. [26] proposed an intelligent system for various energy policies in response to high energy consumption in datacenters. In the case of a high utilization rate, if appropriate policies are adopted, energy consumption of about 20% can be saved without the performance of the datacenter experiencing a significant impact. Nosayba et al. [27] provided a multi-faceted study of temperature management in datacenters. They used a large collection of field data from different production environments to study the impact of temperature on hardware reliability, including the reliability of the storage subsystem, the memory subsystem and server reliability as a whole. Sharma et al. [28] presented a workload distribution method to make the temperature distribution of the datacenter more uniform. They used computational fluid dynamics (CFD) models to evaluate the effectiveness of thermal policies through fault injection simulations and computational dynamics of load calculations. Kien et al. [29] studied the impact of load placement policies on cooling and maximum datacenter temperatures in cloud service providers that operate multiple geographically distributed datacenters. Then, they proposed dynamic load distribution policies that consider all electricity-related costs as well as transient cooling effects. Above all, it can be seen that some energy-aware methods can also improve the energy efficiency to a great extent.

2.4. Renewable Energy Utilization in Datacenters

Nowadays, some IT companies operating large-scale datacenter are considering using green energy as part of their energy supply. Therefore, there are some studies have researched how to schedule loads in the datacenter and, in turn, make full use of green energy. Li et al. [30] proposed a lightweight server power management scheme that maintains application performance by switching between wind energy and traditional power. Goiri et al. [31] proposed the *GreenSlot* framework for scheduling tasks and the *GreenHadoop* framework for MapReduce type tasks [32]. These proposals are based on the prediction of the availability of renewable energy, through different scheduling strategies to maximize the use of green energy. Wang et al. [33] also focused on green datacenters, using solar energy as part of the energy supply in datacenters, and proposed a green-energy-aware virtual machine migration strategy to maximize the use of solar energy. There are some studies that proposed to schedule or allocate tasks by forecasting renewable energy in advance. Baris et al. [34] designed an adaptive datacenter tasks scheduler to maximize the use of renewable energy. They utilize short-term prediction of solar and wind energy production to scale the number of tasks according to the expected energy availability. Grange et al. [35] presented an approach for scheduling batch jobs with due date constraints, which takes into account the availability of the renewable energy to reduce the need of brown energy and, therefore, running cost. The approach they proposed differs from the existing methods by providing a scheduling algorithm agnostic of the electrical infrastructure. A separate system, managing the renewable sources, provides an arbitrary objective function, which is used to guide the scheduling heuristic. Courchelle et al. [36] have studied both the storage and utilization of photovoltaic energy. They detailed their design of a scheduler that uses solar energy production to make an off-line decision. This enables the virtual machine to be dispatched to the datacenter by the proposed different algorithms, thereby reducing brown energy consumption. These researches did not consider the comprehensive energy factor of datacenter, some of these studies only consider one type of workload, and others do not consider the high energy consumption of datacenter cooling equipment.

2.5. Simulation Tools Comparison

In recent years, researchers have proposed a variety of simulation tools to solve different problems for cloud computing. Dzmitry et al. [37] presented a simulation environment for energy-aware cloud computing datacenters, where the simulation tool *GreenCloud* is a packet-level simulator. Damián et al. [38] proposed the simulation tool *SCORE*, which is dedicated to the simulation of energy-efficient monolithic and parallel-scheduling models and for the execution of heterogeneous, realistic, and synthetic workloads. They also presented a different simulation tool for cloud computing, *GAME-SCORE*, which implements a scheduling model based on the *Stackelberg* game [39]. Leandro et al. [40] proposed *CoolEmAll* project, which is used for modeling and simulating energy-efficient and thermal-aware datacenters. The aim of *CoolEmAll* is to address energy-thermal efficiency of datacenters by combining the optimization of IT, cooling, and workload management. Due to the fact that the simulation tool *Cloudsim* proposed in Reference [41] is an open source tool, it is more suitable for simulating the problem solved in our paper. Therefore, we chose the *Cloudsim* tool to simulate the method we proposed.

Compared with the prior work, the main contributions of our work include:

- This paper mainly takes into account two types of hybrid workload running upon the datacenter, including interactive workloads and batch-type workloads. A part of the batch-type workloads will be deferred to be executed until renewable energy is sufficient. In this way, the variable energy input can be more efficiently utilized for task processing.
- During the workload management procedure, we also jointly considered the non-IT energy consumption by adjusting the supply air temperature of the cooling device. We take pre-cooling actions when there is surplus solar energy, and adjust the temperature setting of the cooling devices dynamically according to the current surplus energy, thus avoiding the possible waste of the extra generated power and responding to the need for cooling when solar energy is insufficient.

3. The Architecture of the Green Datacenter

In this section, we present the architecture of the green datacenter using mixed energy supplies and the prediction model used to forecast the energy generation amount.

3.1. Datacenter Architecture

Assume the datacenter system consists of *N* hosts, denoted as host 1 to host *N*. These hosts complete distributed tasks individually or collaboratively. The resources of the host generally include CPU, storage, bandwidth (*bw*). We use c_{max} to represent the maximum CPU capacity of a host, the maximum storage and the maximum *bw* of a host can offer is denoted as s_{max} and b_{max}, respectively. We use $R(c, s, b)$ to represent the total resources of a host. The total simulation time is one day and divided into τ = 24 time slots, and thus the length of each time slot is 60 min. In order to schedule the tasks and adjust the temperature provided in each time slot, we assume that the supply air temperature of the cooling device can be set dynamically on demand.

Figure 1 depicts the architecture of the green datacenter powered by both renewable energy and traditional energy from the utility grid. The grid utility and renewable energy are combined together through the automatic transfer switch in order to provide power for the datacenter. The IT devices include servers, storage, and networking switches that support applications and services hosted in the datacenter. The cooling devices deliver the cooling resources to dispatch the heat generated by IT equipment. In this paper, the cooling capacity is delivered to the datacenter through the computer room air conditioning units (*CRAC*) from the cooling micro-grid that consists of the traditional chiller plant. The architecture does not consider the energy storage equipment, because the energy storage equipment has the following shortcomings [32]: (a) The internal resistance and self-discharge of the battery can result in loss of energy; (b) the battery-related costs predominate in solar powered systems; and (c) the chemicals in the battery can cause harm to the environment to some extent.
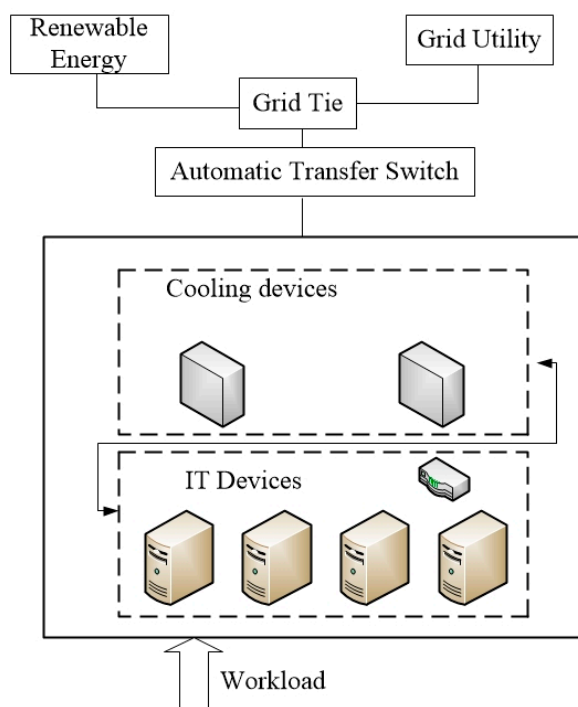
Figure 1. Green datacenter architecture.

3.2. Renewable Energy Forecasting

Considering the use of renewable energy as a part of the energy supply in the sustainable datacenter and the fact that the solar power generation is unknown and unstable, we conducted a prediction of the amount of solar energy generated in advance to better schedule the batch-type jobs. In this way, the impact of unstable solar energy supply on datacenter scheduling jobs can be avoided to some extent. There are many researchers who use a variety of methods to predict the amount of power generation [11,31]. In this paper, we use solar energy as the renewable energy supply. The neural network model LSTM (long short-term memory) is adopted for solar prediction, which is a recurrent neural network trained using backpropagation through time and overcomes the vanishing gradient problem. Because the LSTM model adds a "processor" to determine whether the information is useful or not, it has a better memory function for historical data than some other neural network models, and this model is suitable for processing and predicting important events with relatively long intervals and delays in the time series. We select *m*-day historical solar power generation data as a training dataset to predict the solar power generation on day *m* + 1.

In order to have a precise prediction to achieve a more accurate allocation of jobs, we derived the data from the public data sharing website [42]. We selected solar power data for sunny days and cloudy (rainy) days in February, March, and April 2018 as training sets to forecast the amount of solar power generation of 14 July 2018 and the amount of solar power generation of 8 May 2018, respectively, where the 14 July 2018 is a sunny day and the 8 May 2018 is a cloudy day. We combined the *scikit-learn* library of machine learning for model training and data normalization. The forecast result of a sunny day shown in Figure 2a,b represents the forecast value under a cloudy (rainy) day, wherein the error rate between the predicted and actual values remains within 7% and 20% on a sunny day and cloudy (rainy) day, respectively.
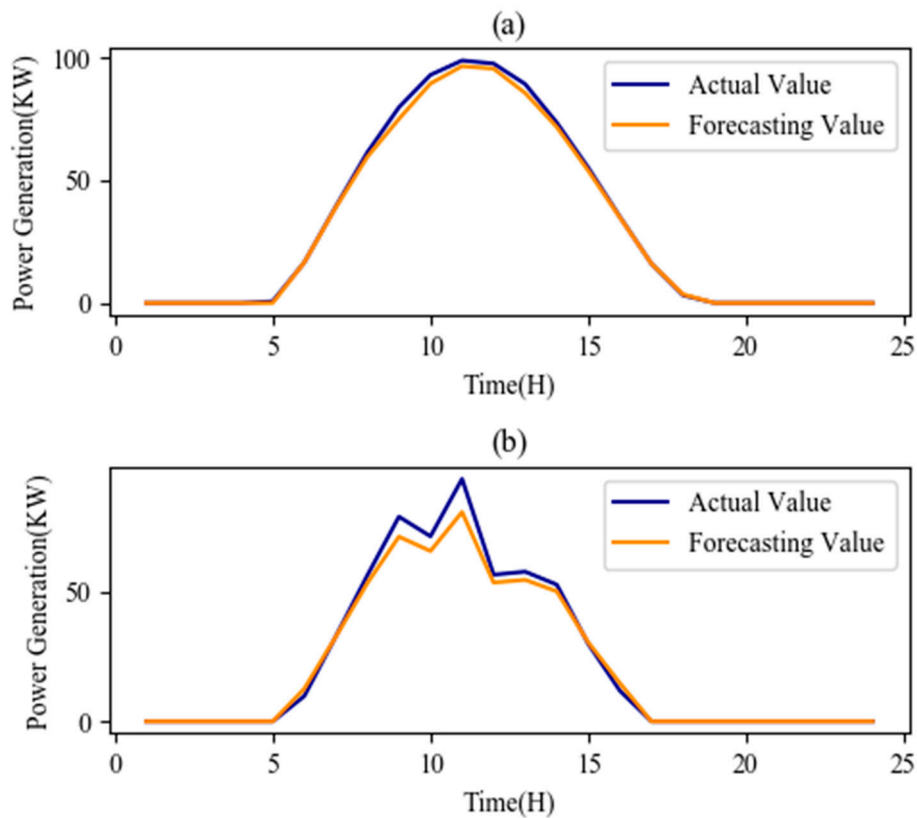
Figure 2. Solar power forecasting: (**a**) a sunny day; (**b**) a cloudy (rainy) day.

4. Problem Statement

4.1. Workload Definition

Here, we consider two main types of IT workloads in the sustainable datacenter: Batch-type and critical interactive type [43,44]. Batch-type workloads are those workloads submitted to a job queue which will be executed when resources become available. The workloads that need to be responded immediately but its execution should be completed before their deadlines are called critical interactive workloads.

To effectively manage the power consumption of the datacenter, we postpone the execution of some time-flexible workloads. Hereafter, we use $I(t)$ and $B(t)$ to represent the amount of the interactive workloads and batch-type workloads submitted at time t, respectively. Let r_i and r_j denote the resources are allocated to an interactive job i and a batch-type job j, respectively. Denote the dt_j as the time of batch-type job j that can be deferred to be processed, and dt_{max} as the maximum time a job that can be delayed. We defer the execution of a batch-type job from time $t \in \tau$ to $t1 \in [t, t + dt_{max}]$. $\varphi(t, t1)$ denotes the number of batch-type jobs submitted at time t that can be deferred to time $t1$. We use $J(t)$ to represent the number of total batch-type workloads, which are needed to be processed at time t. Obviously, the $J(t)$ included the batch-type workloads postponed from previous time slots and a part of the workloads submitted at time t.

$$J(t) = B(t) - \varphi(t, t1) + \sum_{t'=1}^{t-1} \varphi(t', t) \tag{1}$$

where $t1 > t$ and $t' < t \leq \tau$. Therefore, the total workloads $\lambda(t)$ that need to be processed at time t should be calculated as:

$$\lambda(t) = I(t) + J(t). \tag{2}$$

4.2. Power Consumption Model

Usually, the computing power consumption is the dominating part of the total power consumption of the datacenter. The operation of the server includes a variety of hardware devices. The power consumption of the host is related to the voltage supply and operating frequency and is greatly affected by the change of the load. Denote $P_i{}^t$ as the computing power consumption of host i at tth time slot. We adopt a linear growth approach to calculate power consumption.

$$P_i{}^t = P_i^{\max} \cdot (c + (1 - c) \cdot u_i) \tag{3}$$

where the $P_i{}^{max}$ represents the maximum power of a host can consume, c is a constant representing the proportion of idle energy consumption, and u_i represents the CPU utilization of a host i.

Hence, the total computing energy consumption PC_t in each time slot can be calculated by

$$PC_t = \sum_{i=1}^{N} P_i^t \tag{4}$$

where $t = 1, 2 \ldots \tau$.

The power consumption of the cooling system is variable and controllable, which also occupies a high proportion of the total power consumption of the datacenter. In this paper, the servers are cooled by traditional air-cooling technology. The power required for cooling is usually related to how much heat is dissipated. A commonly used measurement coefficient is called *CoP* (coefficient of performance), which is generally defined as the ratio of the heat to be dissipated to the power consumption of the cooling equipment. However, the *CoP* is directly affected by the cooling temperature. Generally, the higher cooling temperature will lead to a higher *CoP* value so that the cooling device itself consumes less power. We adopt the *CoP* model in Reference [18], which is obtained from water chilled *CRAC* (computer room air conditioner) unit in *HP* Utility datacenter, as follows:

$$CoP(T_{\sup}) = 0.0068 T_{\sup}{}^2 + 0.0008 T_{\sup} + 0.458 \tag{5}$$

where T_{sup} is the supply air temperature of the cooling device. Then, the cooling energy consumption PAC_t can be calculated by the following equation:

$$PAC_t = \frac{PC_t}{CoP(T_{\sup})}. \tag{6}$$

The total energy consumption PT_t in a time slot of a datacenter can be calculated as follows:

$$PT_t = PC_t + PAC_t. \tag{7}$$

4.3. Power Management Problem

Since the target datacenter is supplied by mixed energy sources, we adjust the tasks of the datacenter to achieve energy-saving effects while making full use of the generated solar energy. In order to reduce the usage of brown energy, we deferred the execution of some delay-tolerant tasks to the time when solar energy is sufficient. We define the amount of solar energy that can be supplied at a moment as S_t, where $t = 1, 2 \ldots \tau$. Usually, the maximum power consumption value of a host is usually defined in its power consumption model. Therefore, the number of hosts N_t that the datacenter can keep active under the supply of solar power at each time slot could be estimated as

$$N_t = \frac{S_t}{P^{\max}} \tag{8}$$

where $P^{\max}$ is the maximum power consumption of the host in the current power model.

4.4. Thermal-Aware Considerations

The power consumption of the servers will make the surrounding environmental temperature increase, due to the heat recirculation. Through *CFD* (computational fluid dynamics) analysis, previous studies [22,45] have proposed a relationship between the power consumption of the host, indicating that the inlet temperature is governed by the following equation:

$$\mathbf{T_{in}} = \mathbf{T_{sup}} + \mathbf{D} \times \mathbf{P} \tag{9}$$

where $\mathbf{T_{in}}$ and $\mathbf{T_{sup}}$ are the vectors of inlet temperature and supplied air temperature, respectively. $\mathbf{D}$ is the heat transfer matrix [46] and $\mathbf{P}$ is the vector of the power consumption of hosts in the datacenter. In order to avoid overheating of the host and maintain a high performance of the host, the maximum inlet temperature is generally limited. Here, we denote $T_{\max}$ as the maximum inlet temperature. Therefore, under the premise of specifying the maximum inlet temperature $T_{\max}$ of the host, we can implement the pre-cooling for job management.

$$\mathbf{T_{in}} \le T_{\max} \tag{10}$$

Therefore, under the premise of specifying the maximum inlet temperature $T_{\max}$ of the host, we use the pre-cooling to calculate the maximum power consumption that the host can consume.

$$\Delta T = T_{\max} - \mathbf{T_{sup}} \tag{11}$$

Overall, the objective of the optimization problem is to maximize the utilization of solar energy in each time slot t, which can be depicted as

$$\text{minimize } |S_t - PT_t| \tag{12}$$

$$\text{subject to } \sum_{i=1}^{I(t)} r_i + \sum_{j=1}^{J(t)} r_j \le R(c_{\max}, s_{\max}, b_{\max}) \tag{13}$$

$$dt_j <= dt_{\max}, \ \mathbf{T_{in}} \le T_{\max} \tag{14}$$

where in Equations (13) and (14) specify the constraints corresponding to the limits of resources, job deadlines, and the room temperature, respectively.

5. Methods and Strategies

To address the issue proposed and defined in Section 4, we proposed a thermal-aware approach for workload and power management of the datacenter, and also implemented three other methods for comparison. We considered the characteristic of different jobs in the datacenter. We mainly took into account two categories, including interactive workloads and batch-type workloads. As previously described, the interactive workloads should be responded immediately, while the batch workloads are delay-tolerant.

5.1. Static Method (ST)

Under this method, the batch-type jobs will be processed when they arrive as soon as possible without any other scheduling actions.

5.2. Load Balancing Distribution over Time (LB)

Under this strategy, the batch-type jobs are scheduled and distributed evenly over multiple time slots, while the interactive tasks will be processed immediately after they arrive.

5.3. Best Effort Strategy (BS)

Under this strategy, a scheduling plan will be made for the submitted batch-type jobs based on the predicted solar power generation amount. The number of active hosts in a time slot constrained by the supply of solar energy can be calculated by using Equation (9). Then, it should be judged whether the current number of interactive jobs could be handled by these hosts. If the number is not enough, brown energy has to be used to power on some extra hosts according to the demand of jobs.

5.4. Thermal-Aware Workload Management (TM)

Under this strategy, we take both workload and temperature adjustment of the datacenter into account simultaneously. The critical component of this method is the time-shifting of batch-type jobs to consume more solar power. On the basis of *BS* method, we further perform pre-cooling actions if there is surplus solar energy at *t*th time slot and there are no jobs need to be processed so as to deal with more jobs when the solar energy generation becomes insufficient in the future; but if there are jobs that have not been processed, then we use the surplus solar energy to perform the jobs first. Specifically, the temperature will be adjusted dynamically according to the amount of surplus solar energy, denoted as E_t as follows,

$$E_t = S_t - PT_t \tag{15}$$

where $t = 1, 2 \ldots \tau$. We can use Equations (4)–(6) and (15) to calculate the temperature, which the cooling devices should supply given the extra energy from solar generation. In this method, the main consideration is to perform as many jobs as possible when solar power is sufficient and to decrease the power consumption of cooling devices. Compared with the first three strategies, this strategy considered power consumption both IT devices and cooling devices. The pseudo codes of the algorithm are shown as follows (Algorithm 1).

Algorithm 1: The process of *TM* method

Input: the number of jobs, solar power generation, T_{max}
Output: job schedule plan
1.　$S_t \leftarrow$ getSolarPrediction()
2.　$N_t \leftarrow$ calculate the number of hosts can be powered by the provided solar power at time slot t
3.　**if** the number of hosts N_t is enough to process $\lambda(t)$ jobs **then**
4.　　　**if** $E_t > 0$ **then**
5.　　　　　perform the pre-cooling action (decrease the temperature of the cooling device)
6.　　　**end if**
7.　**else**
8.　　　power on some hosts using brown energy according to the workload demand
9.　　　**if** $\mathbf{T_{in}} < =T_{max}$ **and** performed pre-cooling **then**
10.　　　　　allocate more batch-type jobs
11.　　　　　update the remaining number of batch-type jobs
12.　　　**end if**
13.　**end if**
14. **return:** job schedule plan

---

6. Evaluation Results

Here, we set up a series of numerical simulation and experiments to evaluate the four methods proposed in Section 5. We simulated a data center using the *CloudSim-plus* tool, a cloud computing simulation tool that was extended based on the 3.0 version of the *CloudSim* tool, using some of the features proposed by JDK 1.8. The *Cloudsim* is developed in the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, at the Computer Science and Software Engineering Department of the University of Melbourne. We simulate the four methods proposed in this paper under a random load.

We specify the T_{max} = 30 °C, which specifies that the datacenter inside temperature should be less than 30 °C [29]. The solar power data are derived from a public solar data sharing website [42].

　　　Due to the particularity of the load model defined in this paper, we did not consider using actual load tracking data, such as Google traces. Therefore, in order to facilitate the simulation but without loss of generality, we randomly generated jobs that arrived over time and assumed that some batch-type jobs would be submitted at middle night and around noon [11]. Then we use the method proposed in Reference [11] to specify that the maximum time of batch-type jobs could be deferred is 12 h, which means dt_{max} = 12. Figure 3 shows the number of arrived interactive and batch-type jobs during each time slot. The number of servers in datacenter and the power consumption parameters are shown in Table 1.
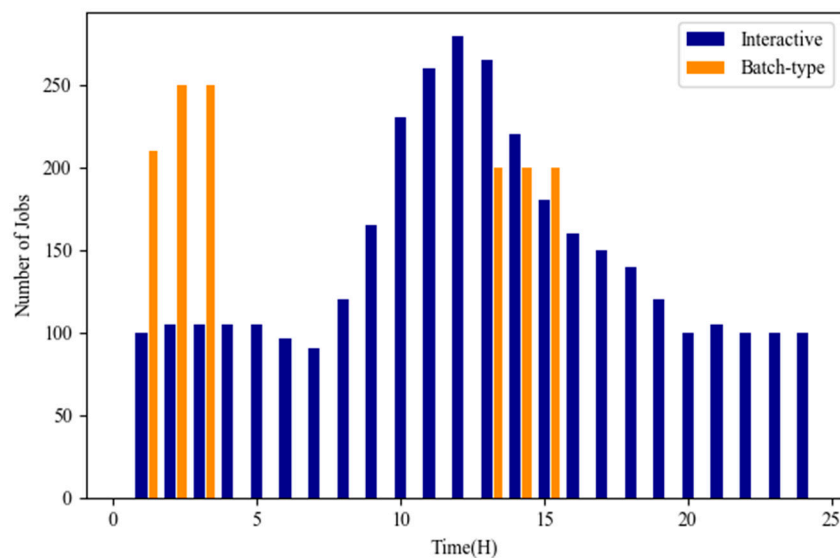

Figure 3. Job arrival numbers.

Table 1. Parameters Settings.

Parameters	P_i^{max}	c	N
Values	100 (w)	0.7	500

6.1. Power Consumption

　　　In this subsection, the results illustrate the solar power utilization in detail under the four strategies described previously. As shown in Figure 4a, the blue column represents solar power usage under different strategies on a sunny day. It can be seen that the solar utilization of *TM* method is the highest, reaching 98%, while *ST* is the lowest. For the sake of having a clear examination, we show detailed utilization values in Table 2. We also illustrate the solar utilization on a cloudy (rainy) day shown in Figure 4b and statistical results in Table 3; we can intuitively obtain that solar power utilization under the *TM* strategy is also the highest in cloudy conditions.

Table 2. Solar utilization percentage under sunny weather condition.

Strategies	*ST*	*LB*	*BS*	*TM*
Solar Utilization Percentage	66.2%	66.8%	89.1%	98.2%

Table 3. Solar utilization percentage under cloudy (rainy) weather condition.

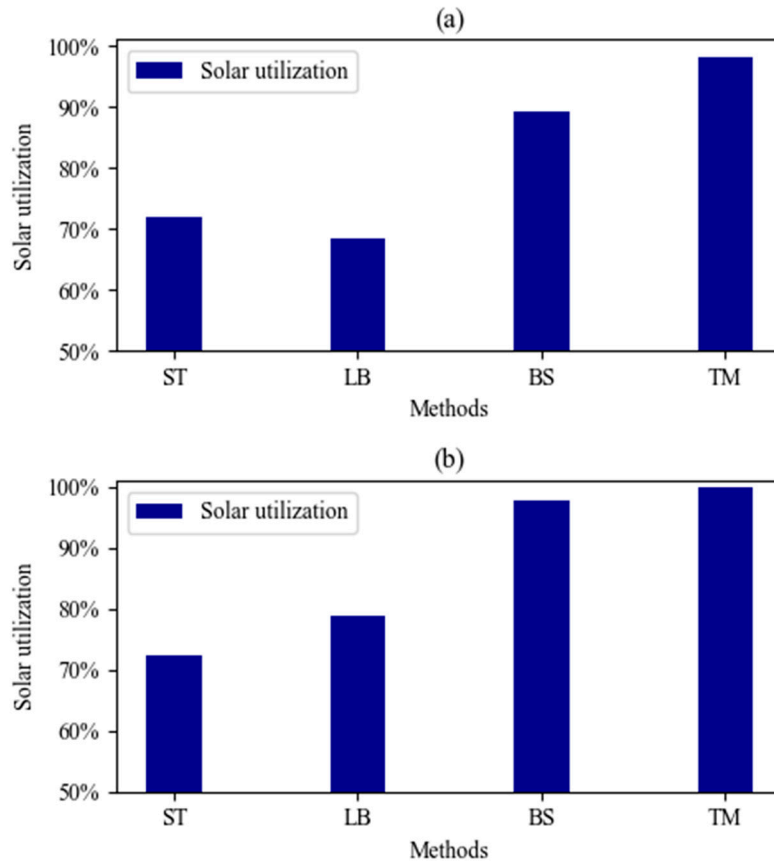| Strategies	*ST*	*LB*	*BS*	*TM*
Solar Utilization Percentage	86.1%	89.2%	96.7%	99.5%

Figure 4. Solar utilization percentage: (**a**) sunny day; (**b**) cloudy (rainy) day.

Figure 5 depicts the detailed power consumption under the four strategies, with power consumption including computing power consumption (batch-type jobs and interactive jobs) and the cooling power consumption. As shown, Figure 5a–d illustrate the particular power consumption under the *ST*, *LB*, *BS*, and *TM*, respectively. Obviously, *ST* and *LB* do not make full use of solar energy but use more of the brown energy for power supply. The power consumption of *BS* can vary according to the supplied solar power, but some solar power is also wasted. Compared with the other three methods, *TM* can make better use of solar energy by jointly scheduling workloads and adjusting the cooling temperature. Hence, *TM* only uses brown energy sources to supply power when necessary.

We also analyzed the detailed power consumption under a cloudy (rainy) day, as shown in Figure 6d, due to the fact that there are some batch-type jobs that need to be processed in addition to interactive jobs when there is solar power available, so there is not surplus solar energy for taking pre-cooling action. This means that the proposed strategy consumes more brown energy on a cloudy (rainy) day, but it also maximizes the use of solar energy. However, the detailed power consumption in the other three methods is similar to the situation in sunny weather conditions.

Here, the results illustrate, in detail, the power consumption and solar energy utilization of the four strategies mentioned previously. As shown in Figure 7, the green column represents the use of solar power under the current strategy and the orange portion represents the use of brown energy used. The total power consumption of *TM* was the most, since it almost consumed all of the generated solar energy, with the least consumption of brown energy. In contrast, the other three methods consumed

more brown energy. For a more detailed explanation, we have listed the solar values actually used under various strategies in Tables 4 and 5 on a sunny and cloudy (rainy) day, respectively. It is clear that the *TM* strategy uses the most solar energy in both weather conditions.
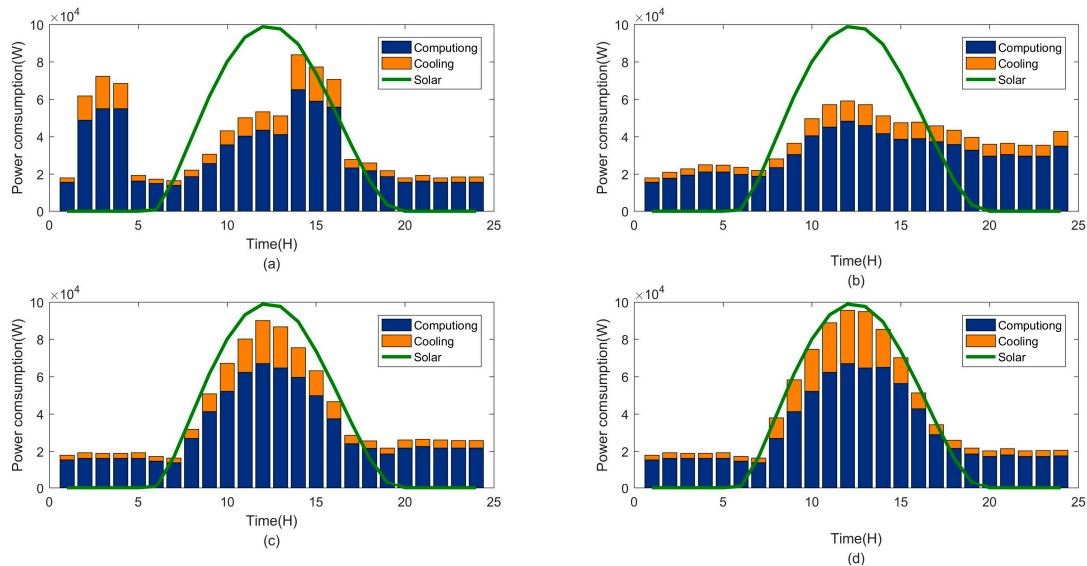

Figure 5. Detailed power consumption under four strategies (sunny): (**a**) *ST*; (**b**) *LB*; (**c**) *BS*; (**d**) *TM*.
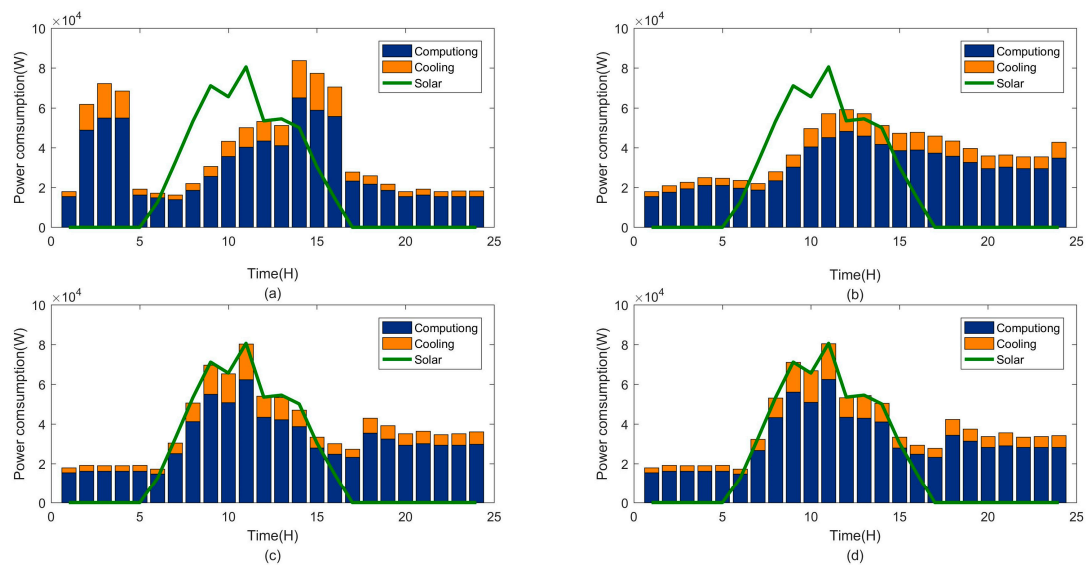

Figure 6. Detailed power consumption under four strategies (cloudy): (**a**) static method (*ST*); (**b**) load balancing distribution over time (*LB*); (**c**) best effort strategy (*BS*); (**d**) thermal-aware workload management (*TM*).

Table 4. Values of actual solar power utilized on a sunny day.

Strategies	ST	LB	BS	TM
Values (KW)	53.1	50.4	65.7	72.4

Table 5. Values of actual solar power utilized on a cloudy (rainy) day.

Strategies	*ST*	*LB*	*BS*	*TM*
Values (KW)	37.4	40.8	50.7	51.7
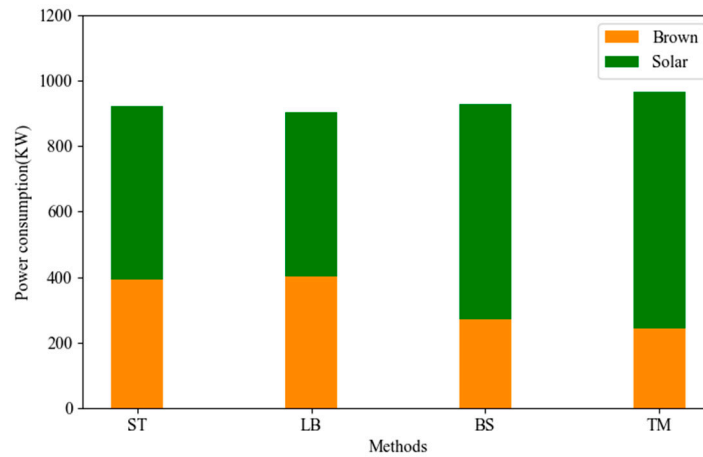
Figure 7. Power usage comparisons on a sunny day.

6.2. Job Scheduling Details

As shown in Table 6, we can obtain the average waiting time that batch-type jobs can be responded to under the four methods. It is obvious that the waiting time under *ST* is the lowest because this method gives the response as quickly as possible to the arrived batch-type jobs. However, under the other three methods, batch-type jobs have different waiting times. The average waiting time under *LB* is longest since some batch-type jobs submitted at morning will be evenly deferred to be executed at all the next time slots. Compared with the *BS* method, the waiting time of the *TM* method is relatively short because more jobs can be performed when the solar energy is insufficient in the afternoon, while the *BS* method will postpone more batch-type jobs to be executed at night, which results in a long waiting time. The average waiting time for batch-type jobs in cloudy (rainy) weather is also given in Table 7. We can see that the waiting time under each method is longer than in sunny weather condition. This is because solar energy is not sufficient in cloudy weather, and the *TM* method does not work well. However, the waiting time of the *TM* method is shorter than other strategies except for the *ST* method. Therefore, we can obtain the *TM* method is suitable for various (Service Level Agreement) SLA-constrained environments, but the energy-saving effect may not be very ideal with strong SLA constraints.

Table 6. Average waiting time of batch-type jobs on a sunny day.

Methods	*ST*	*LB*	*BS*	*TM*
Average Waiting Time (hour)	0	8.3	5.2	4.8

Table 7. Average waiting time of batch-type jobs on a cloudy (rainy) day.

Methods	*ST*	*LB*	*BS*	*TM*
Average Waiting Time (hour)	0	8.3	6.4	5.9

Figure 8 shows the job scheduling conditions under the four methods. We can see from the figure that more jobs can be executed when there is sufficient solar power supply under *TM*. Compared with *BS*, there are more jobs were scheduled than *BS* in several time slots after 12:00. This is because *TM* conducted pre-cooling actions, which made the room cooler when solar power was surplus, and thus

facilitated the later scheduling of more jobs when the solar generation dropped. Furthermore, since a part of batch-type jobs were processed between 13:00 and 16:00, so fewer jobs were processed than *BS* in several time slots after 20:00, which further saved the consumption of the utility grid power.
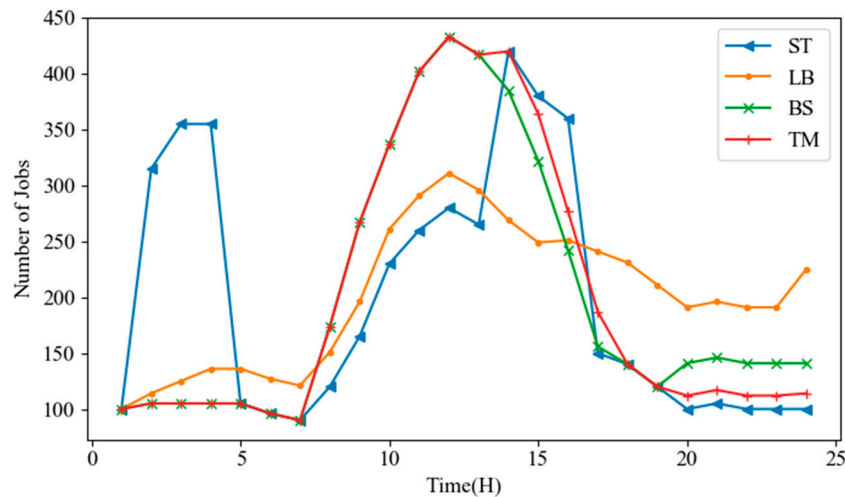

Figure 8. Comparison of job scheduling under the four policies.

6.3. Temperature Provided by the Cooling Device

Figure 9 shows the temperature variation provided by the cooling equipment under *BS* and *TM* strategies. As shown in Figure 9, we can observe that the temperature of *TM* was lower than *BS* in several time slots before 12:00 because the surplus solar energy was not fully utilized and there were no jobs that needed to be processed at these moments, so *TM* used the extra solar energy for cooling. *TM* could fully consider the extra solar power and carry out pre-cooling actions to cope with the power consumption demand for a period of time in the future, thereby utilizing the solar energy more.
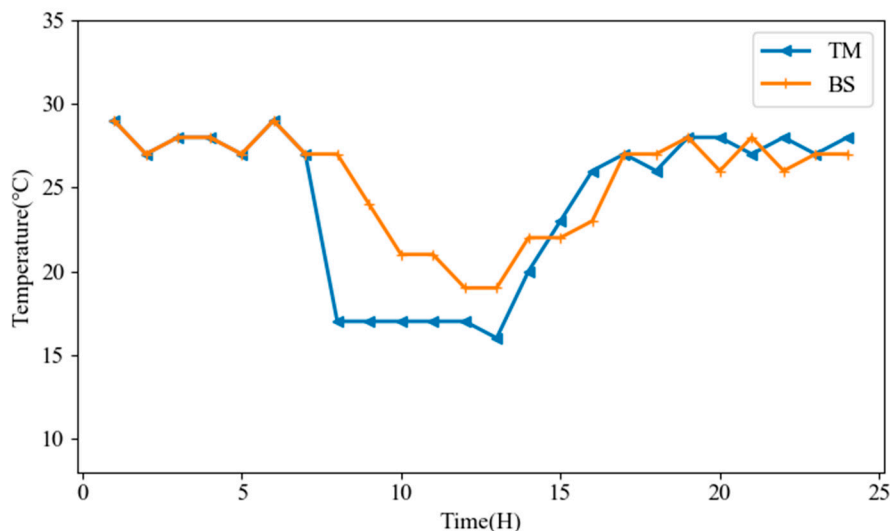

Figure 9. Supply air temperature variation provided by the cooling device.

7. Conclusions and Future Work

In this paper, we studied renewable-aware and thermal-aware workload management approaches to fully use the green energy provided for the datacenter and in turn to minimize the total energy cost. Due to the uncertain nature of renewable energy generation, we use the neural network model to predict solar energy. After considering the multiple characteristics of workloads and high energy cost

of the cooling device, the *TM* strategy proposed in this paper shows a good effect for scheduling hybrid types of workloads and adjusting the temperature dynamically according to surplus solar energy supplied in each time slot in the datacenter. However, the *TM* strategy does not work well when the solar energy supply is insufficient in cloudy (rainy) weather through experimental analysis. But the experiment results illustrate that *TM* can better achieve the energy saving goal as well as minimize the overall power cost of the datacenter no matter on a sunny or cloudy weather condition.

The method proposed in this paper mainly takes into account two types of hybrid workload including interactive workloads and batch-type workloads. By delaying the execution of some more delay-tolerant jobs until the solar energy is sufficient and using excess solar energy for pre-cooling to cope with the cooling needs for the next period of time, the possible waste of the extra generated solar power can be avoided, and more jobs can be scheduled when solar energy is supplied.

Currently, our proposed workload scheduling method only considers the amount of renewable energy generated, in order to help the datacenter maximize the use of solar energy, by scheduling some batch-type workload and adjusting the supply temperature of the cooling equipment. In the future, we plan to combine the more demand response signals on the grid side to enable the datacenter to participate in the response plan by adjusting its load and power consumption.

Author Contributions: Conceptualization X.W. and Y.L.; methodology, Q.P.; software, Y.L.; validation, X.W., P.L. and Q.P.; formal analysis, Q.P.; investigation, P.L.; resources, P.L.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, X.W.; visualization, X.W.; supervision, X.W.; project administration, X.W.; funding acquisition, X.W.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chaudhry, M.T.; Ling, T.C.; Hussain, S.A.; Xin, L.U. Thermal-aware relocation of servers in green data centers. *Front. Inf. Technol. Electron. Eng.* **2015**, *16*, 119–134. [CrossRef]

2. Datacenter Energy Consumption and Efficiency Issues. Available online: http://www.jifang360.com/news/2018731/n3530105256.html (accessed on 20 March 2019).

3. Chen, G.; He, W.; Liu, J.; Nath, S.; Rigas, L.; Xiao, L.; Zhao, F. Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services. *Netw. Syst. Des. Implement.* **2008**, *8*, 337–350.

4. Farahnakian, F.; Liljeberg, P.; Plosila, J. Energy-Efficient Virtual Machines Consolidation in Cloud Data Centers Using Reinforcement Learning. In Proceedings of the 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, Turin, Italy, 12–14 February 2014. [CrossRef]

5. Shi, W.; Liu, Z. Virtual Machine Migration Algorithm to Reduce Energy Consumption in Datacenter, Computer, and digital engineering. *Comput. Digit. Eng.* **2018**, *46*, 39–41. (In Chinese) [CrossRef]

6. Berral, J.L.; Goiri, Í.; Nou, R.; Julià, F.; Guitart, J.; Gavaldà, R.; Torres, J. Towards energy-aware scheduling in data centers using machine learning. In Proceedings of the e-Energy International Conference on Energy-Efficient Computing and Networking, Passau, Germany, 13–15 April 2010. [CrossRef]

7. Katsak, W.; Le, K.; Nguyen, T.D.; Bianchini, R. Parasol and GreenSwitch: Managing datacenters powered by renewable energy. In Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages & Operating Systems, Houston, TX, USA, 16–20 March 2013. [CrossRef]

8. Goiri, I.; Katsak, W.; Le, K.; Nguyen, T.D.; Bianchini, R. Designing and managing data centers powered by renewable energy. *IEEE Micro* **2014**, *34*, 8–16. [CrossRef]

9. Deng, W.; Liu, F.; Jin, H.; Li, B.; Li, D. Harnessing renewable energy in cloud datacenters: Opportunities and challenges. *Netw. IEEE* **2014**, *28*, 48–55. [CrossRef]

10. Kong, F.; Liu, X. A Survey on Green-Energy-Aware Power Management for Datacenters. *ACM Comput. Surv.* **2014**, *47*, 1–38. [CrossRef]

11. Liu, Z.; Chen, Y.; Bash, C.; Wierman, A.; Gmach, D.; Wang, Z. Renewable and cooling aware workload management for sustainable data centers. In Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems, London, UK, 11–15 June 2012. [CrossRef]

12. Li, Y.; Wang, X.; Luo, P.; Yang, X. Temperature-aware Workload Management for Sustainable Datacenters Powered by Renewable Energy. In Proceedings of the 2019 3rd International Conference on High Performance Compilation, Computing and Communications (HP3C-2019), Xi'an, China, 8–10 March 2019.

13. Liu, Y.; Yang, E.; Xu, J. Energy Management for Cloud datacenters. *Telecommun. Sci.* **2012**, *28*, 96–102. (In Chinese) [CrossRef]

14. Zhang, G.; Wan, J.; Zhang, R. Energy consumption minimization algorithm for datacenter average server temperature constraint. *J. Comput. Appl.* **2017**, *37*, 54–57. (In Chinese)

15. Salma, K.; Girish, L. Meta Heuristic Approach for Task Scheduling In Cloud Datacenter for Optimum Performance. *IJARCET* **2015**, *4*, 2070–2074.

16. Pinheiro, E.; Bianchini, R.; Carrera, E. Load balancing and unbalancing for power and performance in cluster-based systems. *Workshop Compil. Oper. Syst. Low Power* **2001**, *180*, 182–195.

17. Yeo, S.; Hossain, M.; Huang, J.C.; Lee, H.H.S. ATAC: Ambient Temperature-Aware Capping for Power Efficient Datacenters. In Proceedings of the 5th ACM Symposium on Cloud Computing, Seattle, WA, USA, 3–5 November 2014. [CrossRef]

18. Wang, L.; Khan, S.U.; Dayal, J. Thermal aware workload placement with task-temperature profiles in a data center. *J. Supercomput.* **2012**, *61*, 780–803. [CrossRef]

19. Luo, J.; Rao, L.; Liu, X. Temporal load balancing with service delay guarantees for data center energy cost optimization. *IEEE Trans. Parallel Distrib. Syst.* **2014**, *25*, 775–784. [CrossRef]

20. Beloglazov, A.; Buyya, R. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurr. Comput. Pract. Exp.* **2012**, *24*, 1397–1420. [CrossRef]

21. Khosravi, A.; Nadjaran Toosi, A.; Buyya, R. Online virtual machine migration for renewable energy usage maximization in geographically distributed cloud datacenters. *Concurr. Comput. Pract. Exp.* **2017**, *29*, e4125. [CrossRef]

22. Islam, M.A.; Ren, S.; Pissinou, N.; Mahmud, A.H.; Vasilakos, A.V. Distributed temperature-aware resource management in virtualized data center. *Sustain. Comput. Inform. Syst.* **2015**, *6*, 3–16. [CrossRef]

23. Urgaonkar, R.; Kozat, U.C.; Igarashi, K.; Neely, M.J. Dynamic resource allocation and power management in virtualized data centers. In Proceedings of the IEEE/IFIP Network Operations and Management Symposium, Osaka, Japan, 19–23 April 2010. [CrossRef]

24. Qouneh, A.; Liu, M.; Li, T. Optimization of resource allocation and energy efficiency in heterogeneous cloud data centers. In Proceedings of the 44th International Conference on IEEE Processing of the In Parallel Processing (ICPP), Beijing, China, 1–4 September 2015. [CrossRef]

25. Fernández-Cerero, D.; Jakóbik, A.; Grzonka, D.; Kołodziej, J.; Fernández-Montes, A. Security supportive energy-aware scheduling and energy policies for cloud environments. *J. Parallel Distrib. Comput.* **2018**, *119*. [CrossRef]

26. Fernández-Cerero, D.; Fernández-Montes, A.; Ortega, J.A. Energy policies for data-center monolithic schedulers. *Expert Syst. Appl.* **2018**. [CrossRef]

27. El-Sayed, N.; Stefanovici, I.A.; Amvrosiadis, G.; Hwang, A.A.; Schroeder, B. Temperature management in data centers: Why some (might) like it hot. In Proceedings of the Acm Sigmetrics/performance Joint International Conference on Measurement & Modeling of Computer Systems, London, UK, 11–15 January 2012. [CrossRef]

28. Sharma, R.K.; Bash, C.E.; Patel, C.D.; Friedrich, R.J.; Chase, J.S. Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Comput.* **2005**, *9*, 42–49. [CrossRef]

29. Le, K.; Bianchini, R.; Zhang, J.; Jaluria, Y.; Meng, J.; Nguyen, T.D. Reducing electricity cost through virtual machine placement in high performance computing clouds. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC 2011, Seattle, WA, USA, 12–18 November 2011. [CrossRef]

30. Li, C.; Qouneh, A.; Li, T. iSwitch: Coordinating and optimizing renewable energy powered server clusters. In Proceedings of the International Symposium on Computer Architecture, Portland, OR, USA, 9–13 June 2012. [CrossRef]

31. Goiri, Í.; Beauchea, R.; Ryan, L.; Kien, N.; Thu, H.; Md, G.; Jordi, T.; Jordi, B. Greenslot: Scheduling energy consumption in green datacenters. In Proceedings of the 24th ACM/IEEE International Supercomputing Conference for High Performance Computing, Networking, Storage and Analysis (SC'11), Seattle, WA, USA, 12–18 November 2011. [CrossRef]

32. Goiri, Í.; Kien, L.; Nguyen, T.D.; Jordi, G. GreenHadoop: Leveraging green energy in data-processing frameworks. In Proceedings of the 7th ACM european conference on Computer Systems, Bern, Switzerland, 10–13 April 2012. [CrossRef]

33. Wang, X.; Zhang, G.; Yang, M.; Zhang, L.A. Green-aware Virtual Machine Migration Strategy for Sustainable Datacenter Powered by Renewable Energy. *Simul. Model. Pract. Theory* **2015**, *58*, 3–14. [CrossRef]

34. Aksanli, B.; Venkatesh, J.; Zhang, L.; Rosing, T. Utilizing free energy prediction to schedule mixed batch and service jobs in data centers. *ACM SIGOPS Oper. Syst. Rev.* **2011**, *45*, 53–57. [CrossRef]

35. Grange, L.; Da Costa, G.; Stolf, P. Green it scheduling for data center powered with renewable energy. *Future Gener. Comput. Syst.* **2018**, *86*. [CrossRef]

36. De Courchelle, I.; Guérout, T.; Da Costa, G.; Monteil, T.; Labit, Y. Green Energy efficient scheduling management. *Simul. Model. Pract. Theory* **2018**, *93*, 208–232. [CrossRef]

37. Kliazovich, D.; Bouvry, P.; Khan, S.U. Greencloud: A packet-level simulator of energy-aware cloud computing data centers. *J. Supercomput.* **2012**, *62*, 1263–1283. [CrossRef]

38. Fernández-Cerero, D.; Fernández-Montes, A.; Jakóbik, A.; Kołodziej, J.; Toro, M. Score: Simulator for cloud optimization of resources and energy consumption. *Simul. Model. Pract. Theory* **2018**, *82*, 160–173. [CrossRef]

39. Fernández-Cerero, D.; Jakóbik, A.; Fernández-Montes, A.; Kołodziej, J. GAME-SCORE: Game-based energy-aware cloud scheduler and simulator for computational clouds. *Simul. Model. Pract. Theory* **2018**, *93*, 3–20. [CrossRef]

40. Cupertino, L.F.; Costa, G.D.; Oleksiak, A.; Piatek, W.; Zilio, T. Energy-efficient, thermal-aware modeling and simulation of datacenters: The coolemall approach and evaluation results. *Ad Hoc Netw.* **2015**, *25*, 535–553. [CrossRef]

41. Calheiros, R.N.; Ranjan, R.; Beloglazov, A.; De Rose, C.A.F.; Buyya, R. Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw. Pract. Exp.* **2011**, *41*, 23–50. [CrossRef]

42. SmartPV. Available online: http://www.lvsedianli.com/ (accessed on 19 March 2019).

43. Singh, S.; Chana, I. QoS-aware autonomic resource management in cloud computing. *ACM Comput. Surv.* **2015**, *48*, 1–46. [CrossRef]

44. Gill, S.S.; Buyya, R. Failure management for reliable cloud computing: A taxonomy, model and future directions. *Comput. Sci. Eng.* **2018**. [CrossRef]

45. Mukherjee, T.; Banerjee, A.; Varsamopoulos, G.; Gupta, S.K.S.; Rungta, S. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Comput. Netw.* **2009**, *53*, 2888–2904. [CrossRef]

46. Tang, Q.; Gupta, S.K.S.; Varsamopoulos, G. Thermal-Aware Task Scheduling for Data centers through Minimizing Heat Recirculation. In Proceedings of the IEEE International Conference on Cluster Computing, Austin, TX, USA, 17–20 September 2007. [CrossRef]