# Energy-Aware Resource Management and Green Energy Use for Large-Scale Datacenters: A Survey

Xiaoying Wang, Xiaojing Liu, Lihua Fan and Jianqiang Huang

**Abstract** As cloud computing gains a lot of attention that provides various service abilities, large-scale datacenters become dominant components of the cloud infrastructure. Huge energy consumption appears to be nonignorable leading to significant cost and also bad impacts on the global environment. How to efficiently manage the services while keeping energy consumption under control is an important problem. According to the analysis of prior representative literature, this paper makes an overview of energy-aware resource management approaches. The basic architecture of cloud datacenters and virtualization technology is introduced. Then, we conduct a survey of energy-aware approaches for adaptive resource management of such cloud environments. We also focused on some studies of renewable energy usage for green datacenters. Finally, the research problems are summarized and analyzed synthetically and possible future directions are given.

**Keywords** Cloud computing · Large-scale datacenters · Renewable energy · Energy-aware resource management

X. Wang (✉) · X. Liu · L. Fan · J. Huang
Department of Computer Technology and Applications,
Qinghai University, Xining 100086 Qinghai, China
e-mail: wxy_cta@qhu.edu.cn

X. Liu
e-mail: liuxj@qhu.edu.cn

L. Fan
e-mail: fanlh@qhu.edu.cn

J. Huang
e-mail: huangjq@qhu.edu.cn

S. Patnaik and X. Li (eds.), *Proceedings of International Conference on Computer Science and Information Technology*, Advances in Intelligent Systems and Computing 255, DOI: 10.1007/978-81-322-1759-6_64,

## 1 Introduction

In the cloud environment, the key infrastructure is usually comprised of large-scale datacenters, providing online computing services for thousands of millions of customers simultaneously. These datacenters own hundreds to thousands of heterogeneous server nodes, which could consume significant amount of energy. Furthermore, the ratio of energy cost can reach to more than 50 % compared with the total operational cost of the entire datacenter [1]. Each single year, more than 30 billion dollars are spent on dealing with the extra heat derived from massive enterprise services all over the world, even more than the money spent on buying new hardware and devices [2]. The problems it brought include not only the expensive maintenance and operational cost for datacenter providers, but also the pollution and bad impact on the global natural environment. One of the main reasons is that the power consumed by large-scale datacenters mainly comes from traditional manners [1], which use fuels and coals to generate power.

In order to reduce the total energy consumption of the entire datacenter by itself, a number of methods have been exploited and tried, including: (1) improvement of the chip manufacturing to reduce the hardware power consumption while keeping high performance; (2) server consolidation using virtualization techniques to reduce the number of active server nodes [3]; (3) rebuilding the heat dissipation systems, e.g., using new novel heat dissipation methods such as water cooling to reduce the power consumption. These approaches are all from similar point of view—saving and reducing. However, the basic function of large-scale datacenters is to provide services for high-performance computing and massive data processing, which would more or less limit the reduction amount of energy consumption. According to such considerations, a number of famous enterprises and IT service companies began to explore possible solutions of using renewable energy to provide the power supply for large-scale datacenters. For example, *Google* has been pondering a "floating datacenter" that could be powered and cooled by the ocean [4]; *Apple* planned to build a brand-new datacenter in Prineville, Oregon, and vowed to use "100 % renewable energy" [5]; *HP* also attempted to create a "Net-Zero" datacenter that requires no net energy from utility power grids [6].

Nevertheless, the generation of renewable energy is usually intermittent. For example, solar energy will be greatly impacted by the strength of the direct sunlight, which is usually high in daytime and low in nighttime. Hence, it is a great challenge to appropriately manage the resources of the datacenter and the workloads of the upper-level applications, in order to accurately control the energy consumption to match the fluctuation of the unstable incoming energy supply.

In this paper, we intend to review the related work about energy-aware resource management and the utilization of renewable energy in large-scale datacenters for cloud computing. The architecture and the execution mechanisms of such datacenters are first introduced, and then, we discuss some possible solutions to save energy while keeping system performance. The difficulties and challenges when leveraging the renewable energy are presented, and we will also review some existing approaches that aim to solve the intermittency and fluctuation features. After the comprehensive survey and relevant analysis, we will discuss about the possible future research directions based on the state of the art.

## 2 Overview of Large-Scale Green Datacenters

### *2.1 Architecture*

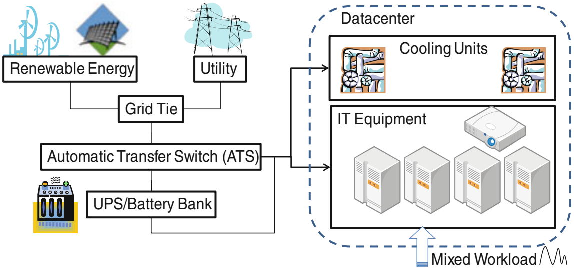
The architecture of a typical large-scale datacenter with mixed energy supplies is shown in Fig. 1. The left half of the figure shows the supplying part of the whole system, which integrates the traditional grid utility and renewable energy. The automatic transfer switch (ATS) combines different energy supplies together and provides the energy to the datacenter. The right half of the figure shows the consumption part of the whole system. The functional equipments inside the datacenter consume energy for dealing with fluctuating incoming workloads. At the same time, some cooling units have to work in order to lower the temperature and guarantee the availability of the devices, which will consume considerable amount of power too.

### *2.2 Virtualization*

Virtualization is an essential technology for cloud computing, which introduces a software abstraction layer between the hardware and the operating system with applications running upon it. Researches in labs and universities are developing approaches based on virtual machines to solve manageability, security, and portability problems [7, 8]. The ability of multiplexing hardware and server consolidation greatly facilitates the requirements of cloud computing.

Specifically, in a large-scale datacenter for cloud computing, virtual machines are usually deployed and used in a manner [9] as the architecture shown in Fig. 2. Multiple VMs can concurrently run applications based on different operating system environments on a single physical server in the datacenter. VMs can be

**Fig. 1** Architecture of the large-scale green datacenter



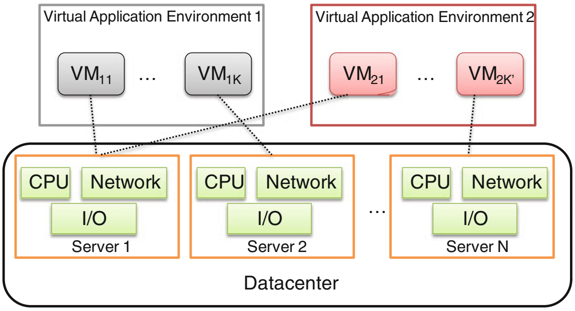


**Fig. 2** Cloud infrastructure with virtual application environments

dynamically started and stopped according to incoming requests, providing flexibility of configuring various partitions of resources on the same physical machine according to different requirements of service requests [10].

# 3 Energy-Aware Resource Management

## *3.1 Switching On/Off Nodes*

A straightforward consideration for saving energy is to properly shutdown some of the physical servers to reduce extra overhead, under a condition of compelling all of the workload of these servers outside. It requires the resource manager to efficiently analyze and redistribute loads and adjust resource allocation schemes. Some researchers have made efforts on such approaches. Chen et al. [11] characterized the unique properties, performance models, and power models of connection servers. Pinheiro et al. [12] developed systems that dynamically turn cluster nodes on and off to handle the load imposed on the system while at the same time save power under light load. In this way, power consumption of the whole system could be saved remarkably by turning some servers into off mode. However, the overhead and time latency of turning on/off nodes is also nonignorable, which might result in a delay in processing of workloads.

## *3.2 Dynamic Voltage and Frequency Scaling*

Comparatively, another more finer-grained way to save energy for CPUs is to utilize the features of dynamic voltage scaling (DVS). Bohrer et al. [13] from *IBM Research* conducted a research on the impact of the workload variation in energy consumption and designed a simulator to quantify the benefits of dynamically scaling the processor voltage and frequency. Rajamony et al. [14], also from *IBM Research*, considered the energy saving issue from two different aspects and proposed independent voltage scaling (IVS) and coordinated voltage scaling

(CVS) policies. Sharma et al. [15] investigated adaptive algorithms for DVS in quality of services (QoS)-enabled web servers to minimize energy consumption subject to service delay constraints. Petrucci et al. [16] presented in their work a dynamic configuration approach for power optimization in virtualized server clusters, which leveraged a dynamic configuration model and outlined an algorithm to dynamically manage the virtual machines, with the purpose of controlling power consumption while meeting the performance requirements.

To sum up, using DVFS techniques to eliminate unnecessary waste for CPU power is a popular way in current researches. Since it supports fast switching with neglectable delay time, the workloads can be processed in time with low overhead.

### *3.3 Saving Cooling Energy Consumption*

The previous subsections presented some existing approaches to reduce the power consumption of IT devices themselves. However, in large-scale datacenters, the energy consumption for heat dissipation and cooling also occupies a significant part of the total amount, even up to 50 % [17]. Hence, some researchers turned to focus on temperature-aware or thermal-aware resource management approaches.

Tang et al. [18] looked into the prospect of assigning the incoming tasks around the data center in such a way so as to make the inlet temperatures as even as possible, allowing for considerable cooling power savings. Pakbaznia et al. [19] presented a power and thermal management framework for datacenters where resources are dynamically provisioned to meet the required workload while ensuring that a maximum temperature threshold is met throughout the datacenter. Ahmad and Vijaykumar [20] proposed *PowerTrade* to trade off idle power and cooling power for each other and reduce the total power; and *SurgeGuard* to over-provision the number of active servers beyond that needed by the current loading so as to absorb future increases in the loading. Wang et al. [21] established an analytical model that describes datacenter resources with heat transfer properties and workloads with thermal features.

To sum up, this section reviews the researches in the area of resource management, task scheduling, load balancing that also partially considers energy consumption reduction. However, since the total energy consumption is significant for such large-scale datacenters, there is finally a limitation of the possible saved energy amount. Even if 10–20 % of the total energy consumption could be saved, the carbon emission will still be a huge amount.

## 4 Renewable Energy Use in Green Datacenters

New types of renewable energy such as solar, wind, and tidal bring advantages by their features including: sufficiency, cleanness, sustainability, nonpollution, and so on. This section summarizes some works about renewable energy use in green datacenters and illustrates some possible attempts.

## *4.1 Single Datacenter*

Here, some researches about how to efficiently utilize renewable energy inside a single datacenter are first reviewed and summarized, as follows.

Deng et al. [22] proposed the concept of carbon-aware cloud applications, which treated carbon-heavy energy as a primary cost, provisioning a cloud instance only if its emission costs are justified by application-specific rules. Goiri et al. designed a framework called *GreenSlot* [23] that aims to schedule batch workloads, and another framework called *GreenHadoop* [24] that orients MapReduce-based tasks. Both are based on the prediction of the availability of renewable energy and try to maximize the utilization of available green energy by different scheduling strategies. Krioukov et al. [25] presented an energy agile cluster that is power proportional and exposes slack. Li et al. [26] proposed *iSwitch*, which switches between wind power and utility grid following renewable power variation characteristics, leverages existing system infrastructures. Arlitt et al. [6] from *HP Labs* introduced and designed a "Net-Zero energy" datacenter managed in a manner that uses on-site renewables to entirely offset the use of any nonrenewable energy from the grid.

## *4.2 Multiple Distributed Datacenters*

Some big companies and enterprises usually establish multiple datacenters around different areas all over the world, powered by mixed green energy and utility grid, as shown in Fig. 3. Recently, there are also some works discussing the possibility of exploring the heterogeneousness of the distributed datacenters and co-scheduling the workload among multiple datacenters.

Stewart and Shen [27] outlined a research agenda for managing renewable in the datacenter, which compliments ongoing efforts to integrate renewables into the grid. Akoush et al. [28] introduced a design called "Free Lunch" that exploits otherwise wasted renewable energy by colocating datacenters with these remote energy sources and connecting them over a dedicated network. Chen et al. [29] proposed a holistic workload scheduling algorithm, called *Min Brown*, to minimize the brown energy consumption across multiple geographically distributed datacenters with renewable energy sources. Le et al. [30] sought to exploit geographically distributed datacenters that pay different and perhaps variable electricity prices, the benefit of different time zones and near sites that produce renewable electricity. Heddeghem et al. [31] looked at the feasibility of globally distributing a number of these renewable sources for powering already distributed datacenters and provided a mathematical model for calculating the carbon footprint. Li et al. [32] proposed a collaborative cost optimization framework by coupling utilities with datacenters via dynamic pricing.

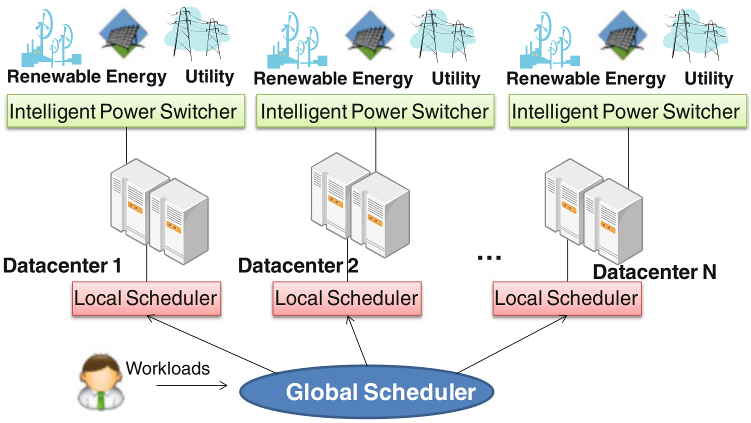


**Fig. 3** Architecture of geographically distributed datacenters

To sum up, the above works considered to utilize the benefit of geographically distributed locations, different time zones, and prices to schedule and dispatch loads onto multiple datacenters.

## 5 Conclusion and Future Directions

In this paper, we reviewed relevant research works about energy-aware resource management approaches inside large-scale datacenters for cloud computing. From the comprehensive survey, we found that although there have been a number of researches starting to explore the energy-efficient management issues, the relevant study of renewable energy usage is still preliminary. Considering that the generation process of renewable energy is usually intermittent and random, we intend to discuss some possible future research directions as follows: (1) It is necessary to study on how to incorporate energy-related metrics into the optimization objectives. (2) A holistic framework has to be established, which could describe the relationships between job scheduling and IT business power consumption, resource allocation and power consumption, transaction and cooling power consumption, and so on. (3) The looseness and slackness of the dominant workload inside the datacenter should be further exploited, which could facilitate the adjustment of resource allocation toward the requirements of varying power supply strength. (4) Since the utilization of all nodes in the datacenter is usually unbalanced, the resource-controlling approach should be aware of temperature and locations. How to design location-aware and thermal-aware strategies is still an important open issue that needs to studied.

**Acknowledgments** This research is funded in part by National Natural Science Foundation of China (No. 61363019, No. 60963005) and Tsinghua—Tencent Joint Laboratory for Internet Innovation Technology (No. 2011-1).

## References

1. Le, K., Bilgir, O., Bianchini, R., et al.: Managing the cost, energy consumption, and carbon footprint of internet services. In: Proceedings of ACM SIGMETRICS Performance Evaluation Review, p. 357–358. ACM (2010)
2. Bianchini, R., Rajamony, R.: Power and energy management for server systems. Computer **37**(11), 68–76 (2004)
3. Uddin, M., Rahman, A.A.: Server consolidation: an approach to make data centers energy efficient and green. Int. J. Sci. Eng. Res. **1**(1), 1–7 (2010)
4. LaMonica, M.: Google files patent for wave-powered floating data center. Available from http://news.cnet.com/8301-11128_3-10034753-54.html (2008)
5. Rogoway, M.: Apple outlines 'green' energy plans for Prineville data center. http://www.oregonlive.com/silicon-forest/index.ssf/2013/03/apple_outlines_green_energy_pl.html (2013)
6. Arlitt, M., Bash, C., Blagodurov, S., et al.: Towards the design and operation of net-zero energy data centers. In: Proceedings of IEEE ITherm2012, pp. 552–561. IEEE (2012)
7. Zhao, M., Figueiredo, R.J.: Experimental study of virtual machine migration in support of reservation of cluster resources. In: Proceedings of Experimental Study of Virtual Machine Migration, pp. 1–8. ACM (2007)
8. Sundararaj, A.I., Sanghi, M., Lange, J.R., et al.: Hardness of approximation and greedy algorithms for the adaptation problem in virtual environments. In: Proceedings of ICAC '06, pp. 291–292. IEEE (2006)
9. He, L., Zou, D., Zhang, Z., et al.: Developing resource consolidation frameworks for moldable virtual machines in clouds. Future Gener. Comput. Syst. (2012). doi: 10.1016/j.future.2012.05.015 (in press)
10. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
11. Chen, G., He, W., Liu, J., et al.: Energy-aware server provisioning and load dispatching for connection-intensive internet services. In: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, pp. 337–350. NSDI (2008)
12. Pinheiro, E., Bianchini, R., Carrera, E.V., et al.: Load balancing and unbalancing for power and performance in cluster-based systems. In: Workshop on Compilers and Operating Systems for Low Power, pp. 182–195. Barcelona, Spain (2001)
13. Bohrer, P., Elnozahy, E., Keller, T., et al.: The case for power management in web servers. Graybill, R., Melhem, R. (eds.) Power Aware Computing. Springer, New York (2002)
14. Rajamony, R., Elnozahy, M., Kistler, M.: Energy-efficient server clusters. In: Proceedings of the Second Workshop on Power Aware Computing Systems. Springer (2002)
15. Sharma, V., Thomas, A., Abdelzaher, T., et al.: Power-aware QoS management in web servers. In: Proceedings of RTSS'03, p. 63. IEEE (2003)
16. Petrucci, V., Loques, O., Niteroi, B., et al.: Dynamic configuration support for power-aware virtualized server clusters. In: Proceedings of 21th Euromicro Conference on Real-Time Systems. Ireland, (2009)
17. Sawyer, R.: Calculating total power requirements for data centers. White Paper, American Power Conversion, (2004)

18. Tang, Q., Gupta, S., Varsamopoulos, G.: Thermal-aware task scheduling for data centers through minimizing heat recirculation. In: Proceedings of IEEE Cluster Computing, pp. 129–138. IEEE (2007)
19. Pakbaznia, E., Ghasemazar, M., Pedram, M.: Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE), pp. 124–129. European Design and Automation Association (2010)
20. Ahmad, F., Vijaykumar, T.: Joint optimization of idle and cooling power in data centers while maintaining response time. In: Proceedings of the Fifteenth Edition of ASPLOS on Architectural Support for Programming Languages And Operating Systems, pp. 243–256. ACM (2010)
21. Wang, L., Khan, S.U., Dayal, J.: Thermal aware workload placement with task-temperature profiles in a data center. J. Supercomput. **61**(3), 780–803 (2012)
22. Deng, N., Stewart, C., Gmach, D., et al.: Policy and mechanism for carbon-aware cloud applications. In: Proceedings of NOMS2012, pp. 590–594. IEEE (2012)
23. Goiri, Í., Beauchea, R., Le, K., et al.: GreenSlot: scheduling energy consumption in green datacenters. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, p. 20. ACM (2011)
24. Goiri, Í., Le, K., Nguyen, T.D., et al.: GreenHadoop: leveraging green energy in data-processing frameworks. In: Proceedings of the 7th ACM European Conference on Computer Systems, pp. 57–70. ACM (2012)
25. Krioukov, A., Alspaugh, S., Mohan, P., et al.: Design and evaluation of an energy agile computing cluster. Technical Report UCB/EECS-2012-13, University of California, Berkeley, (2012)
26. Li, C., Qouneh, A., Li, T.: iswitch: coordinating and optimizing renewable energy powered server clusters. In: Proceedings of 39th Annual International Symposium on Computer Architecture (ISCA), pp. 512–523. IEEE (2012)
27. Stewart, C., Shen, K.: Some joules are more precious than others: managing renewable energy in the datacenter. In: Workshop on Power Aware Computing and Systems, 2009
28. Akoush, S., Sohan, R., Rice, A., et al.: Free lunch: exploiting renewable energy for computing. In: Proceedings of HotOS 2011, pp. 17–17. USENIX (2011)
29. Chen, C., He, B., Tang, X.: Green-aware workload scheduling in geographically distributed data centers. In: Proceedings of CloudCom 2012, pp. 82–89. IEEE (2012)
30. Le, K., Bianchini, R., Martonosi, M., et al.: Cost-and energy-aware load distribution across data centers. In: Proceedings of HotPower (2009)
31. Van Heddeghem, W., Vereecken, W., Colle, D., et al.: Distributed computing for carbon footprint reduction by exploiting low-footprint energy availability. Future Gener. Comput. Syst. **28**(2), 405–414 (2012)
32. Li, Y., Chiu, D., Liu, C., et al.: Towards dynamic pricing-based collaborative optimizations for green data centers. In: Second International Workshop on Data Management in the Cloud (DMC), pp. 272–278. IEEE (2013)