Executive UPDATE
CUTTER CONSORTIUM
Access to the Experts
BUSINESS TECHNOLOGY & DIGITAL TRANSFORMATION STRATEGIES

Sustainable Cloud Computing:

Foundations and Future Directions

by Rajkumar Buyya and Sukhpal Singh Gill

Major cloud providers such as Microsoft, Google, Facebook, and Amazon rely heavily on data centers to support the ever-increasing demand for their computational and application services. However, the financial and carbon footprint–related costs of running such large infrastructures negatively impact the sustainability of cloud services. Most existing efforts primarily focus on minimizing the energy consumption of servers. In this *Executive Update*, we devise a conceptual model and practical design guidelines for the holistic management of all resources (e.g., servers, networks, storage, cooling systems) to improve energy efficiency and to reduce the carbon footprints in cloud data centers (CDCs). Furthermore, we discuss the intertwined relationship between energy and reliability for sustainable cloud computing, where we highlight the associated issues. Finally, we propose a set of future areas to investigate in the field and propose further practical developments.

CDCs and the Challenge of Sustainable Energy

The cloud computing paradigm offers on-demand, subscription-oriented services over the Internet to host applications and process user workloads. To ensure the availability and reliability of the services, the components of CDCs such as network devices, storage devices, and servers should be run 24/7. Large amounts of data are created by digital activities such as data streaming, file sharing, Internet searching, social networking websites, e-commerce, and sensor networks, and that data can be stored as well as processed efficiently using CDCs. However, creating, processing, and storing each bit of data adds to the energy cost, increases carbon footprints, and further impacts the environment. Due to the large consumption of electricity by CDCs, the community faces the challenge of a sustainable energy economy. Indeed, the amount of energy consumed by CDCs is increasing regularly (see Figure 1) and is expected to be 8,000 terawatt hours (TWh) by 2030.

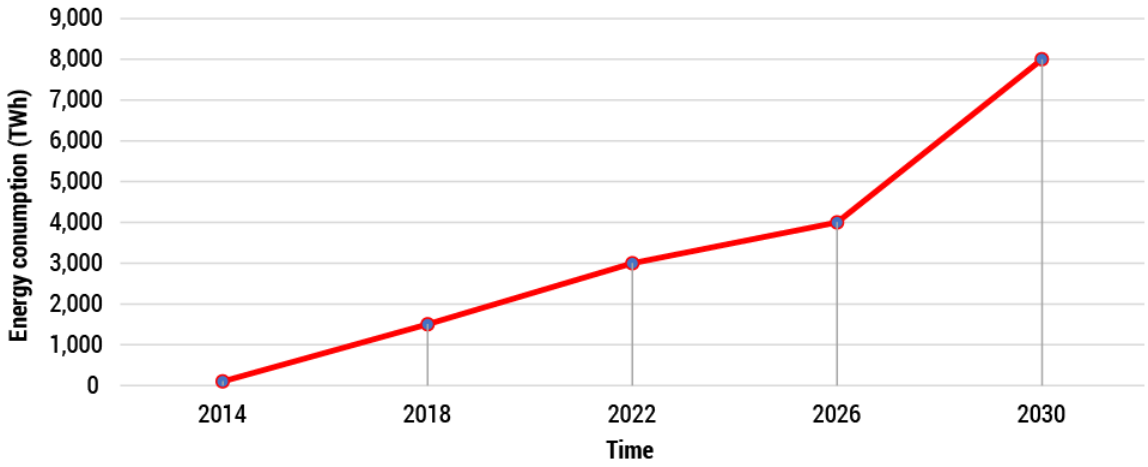

Figure 1 — Energy consumption in cloud data centers. (Source: Andrae and Edler.)

Moreover, existing energy-aware techniques mainly focus on reducing the energy consumption of the servers. Yet other components (e.g., networks, storage, memory, and cooling systems) of CDCs are consuming huge amounts of energy. To improve CDC energy efficiency, there is a need for energy-aware resource management techniques for managing all the resources (i.e., servers, networks, storage, memory, and cooling systems) in a holistic manner. Due to the underloading/overloading of infrastructure resources, the energy consumption in CDCs is not efficient; in fact, most of the energy is consumed while some resources (i.e., networks, storage, memory, processor) sit in an idle state, increasing the overall cost of cloud services.

In the current scenario, CDC service providers are finding alternative ways to reduce the carbon footprint of their infrastructure. Indeed, prominent cloud providers are hoping to power their data centers using renewable energy sources. Future CDCs are required to provide cloud services with a minimum carbon footprint and minimum heat release in the form of greenhouse gas emissions. So what are some of the concerning issues?

Well, an efficient cooling mechanism is required to maintain the temperature of data centers, but it increases costs. Cooling expenses can be decreased by developing waste heat utilization and free cooling mechanisms. But location-aware ideal climatic conditions are needed for an efficient implementation of free cooling and renewable energy production techniques. Moreover, waste heat recovery locations must be identified for an efficient implantation of waste heat recovery prospects. To enable sustainable cloud computing, data centers can be relocated based on: (1) opportunities for waste heat recovery, (2) accessibility of green resources, and (3) proximity of free cooling resources. To resolve these issues and substantially reduce the energy consumption of CDCs, there is a need for cloud computing architectures that can provide sustainable cloud services through a holistic management of resources.

A Conceptual Model

Figure 2 shows a conceptual model for sustainable cloud computing in the form of a layered architecture, which offers holistic management of cloud computing resources to make cloud services more energy-efficient and sustainable. The four main components of the proposed architecture are:

1. **Cloud architecture.** This component is divided into three different subcomponents: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS):

    - *SaaS.* At this layer, an application manager is deployed to handle the incoming user workloads (which can be interactive or batch style) and the transfer to a workload manager for resource provisioning.

    - *PaaS.* At this layer, a controller or middleware is deployed to control the important aspects of the system. An IT device manager covers all the devices attached to the cloud data center. A workload manager controls the incoming workloads from the application manager and identifies the quality of service (QoS) requirement for every workload for successful execution and transfers the QoS information about the workload to the virtual machine (VM)/resource manager. An energy controller manages the energy consumption of the CDC to ensure sustainability of cloud services. A remote CDC manager handles the VM migration and workload migration between local and remote CDCs

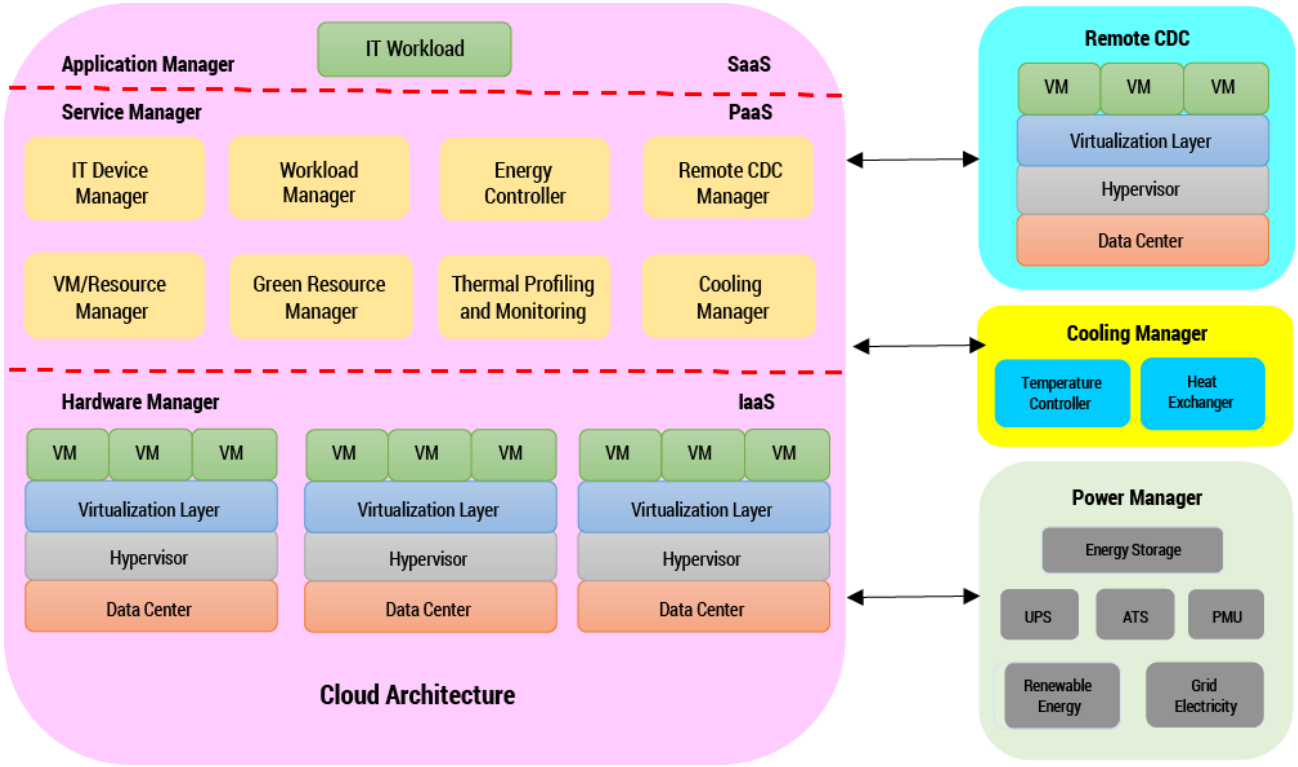

Figure 2 — A conceptual model for sustainable cloud computing.

for effective utilization of energy. A VM/resource manager provisions and schedules the cloud resources for workload execution based on QoS requirements of the workload using physical machines or VMs. A green resource manager, which manages the electricity coming from the power manager, prefers renewable energy as compared to grid electricity to enable a sustainable cloud environment. A thermal profiling and monitoring technique analyzes the temperature variations of the CDC based on the temperature value as monitored by thermal sensors. Finally, a cooling manager controls the temperature of the CDC at the infrastructure level.

- *IaaS.* This layer contains the information about cloud data centers and virtual machines. VM migrations are performed to balance the load at the virtualization layer for efficient execution of workloads. The proactive temperature-aware scheduler monitors the temperature variation of different VMs running at different cores. The power management unit (PMU) is integrated to power all the hardware executing the VMs. Dynamic random-access memory stores the current states of the VMs. A thermal sensor monitors the temperature value; it generates an alert if the temperature is higher than its threshold value and passes the message to the heat controller for further action.

2. **Cooling manager.** Thermal alerts will be generated if the temperature is higher than the threshold value, and the heat controller will act to control the temperature with minimal impact on the performance of the CDC. Electricity coming from an uninterruptible power supply (UPS) is used to run the cooling devices to control the temperature. District heating management, in which the temperature is controlled by using a chiller plant, outside air economizer, and water economizer, is integrated.

3. **Power manager.** The power manager controls the power generated from renewable energy resources and fossil fuels (i.e., grid electricity). To enable a sustainable cloud environment, renewable energy is preferred as compared to grid energy. If there is execution of deadline-oriented workloads, then grid energy can help maintain the reliability of cloud services. The sources of renewable energy are solar and wind. Batteries store the renewable energy. An automatic transfer switch (ATS) manages the energy coming from both sources (i.e., renewable energy and grid electricity) and forwards it to the UPS. A power distribution unit transfers the electricity to all the CDCs and cooling devices.

4. **Remote CDC.** VMs and workloads can be migrated to a remote CDC to balance the load effectively.

Implication of Reliability on Sustainability

Sustainable cloud services are attracting more cloud customers and making cloud more profitable. Improving energy utilization, which reduces electricity bills and operational costs, enables sustainable

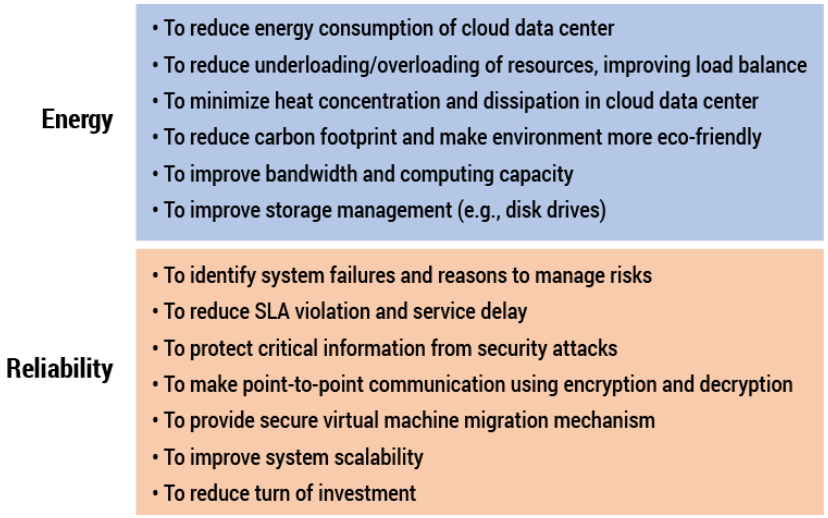

Figure 3 — Issues related to energy and reliability.

cloud computing. On the other hand, to provide reliable cloud services, the business operations of different cloud providers are replicating services, which requires additional resources and increases energy consumption. To overcome this impact, a tradeoff between energy consumption and reliability is required to provide cost-efficient cloud services.

Figure 3 shows the issues related to energy and reliability for sustainable cloud computing. Existing energy-efficient resource management techniques consume a huge amount of energy while executing workloads, which decreases resources leased from CDCs. Dynamic voltage and frequency scaling (DVFS)–based energy management techniques reduce energy consumption, but response time and service delay are increased due to the switching of resources between high-scaling and low-scaling modes. Furthermore, reliability of the system component is also affected by the excessive turning on and off of servers. Power modulation decreases the reliability of server components, such as storage devices, memory, and so on. By reducing energy consumption of CDCs, we can improve the resource utilization, reliability, and performance of the server. Therefore, there is a need for new energy-aware resource management techniques to reduce power consumption without affecting the reliability of cloud services.

Areas to Investigate

The ever-increasing demand for cloud computing services deployed across multiple CDCs necessitates a significant amount of power, resulting in high carbon emissions and a negative effect on the environment. In sustainable cloud computing, renewable energy resources power the cloud data centers, replacing the conventional fossil fuel–based grid electricity or brown energy to effectively reduce carbon emissions. Employing energy-efficient mechanisms also makes cloud computing sustainable by greatly reducing carbon footprints. Waste heat utilization from heat dissipated through servers and employing mechanisms for free cooling of the servers make the CDCs sustainable.

Sustainable cloud computing covers the following elements in making the data center sustainable: (1) using renewable energy instead of grid energy generated from fossil fuels, (2) utilizing the waste heat generated from heat-dissipating servers, (3) using free cooling mechanisms, and (4) using energy-efficient mechanisms. All these factors contribute in reducing carbon footprints, the operational cost, and energy consumption. Issues related to sustainable cloud computing can be organized into seven categories: the application model, virtualization, waste heat utilization, thermal-aware scheduling, renewable energy, resources targeted in energy management, and capacity planning (see Figure 4).

Although industry and academia have explored issues in sustainable cloud computing, there exist many open issues in the context of models for application composition, scheduling, resources targeted in energy management, harnessing of renewable energy and heat generated by resources, and capacity planning.

Application Model

In sustainable cloud computing, the application model plays a vital role. The efficient structure of an application can improve the energy efficiency of CDCs. Application models can be data parallel, function

Figure 4 — Issues in sustainable cloud computing.

parallel, and message passing. The *data parallel model* is a form of parallelization across multiple processors in parallel computing environments. It focuses on distributing the data across different nodes, which operate on the data in parallel. Examples of data parallel models are MapReduce, Bag-of-Task, and parameter sweeps. The *function parallel model* is a form of parallelization of computer code across multiple processors in parallel computing environments. It focuses on distributing tasks concurrently, which are performed by processes or threads across different processors. Examples of a data parallel model are threads and tasks. The *message passing* interface provides a communication functionality between a set of processes, which are mapped to nodes or servers in a language-independent way, encouraging the development of portable and scalable large-scale parallel applications.

Virtualization

During the execution of workloads, *VM migration* is required to balance the load effectively to utilize renewable energy resources in decentralized CDCs. Due to lack of onsite renewable energy, VM techniques migrate workloads to other machines distributed geographically. VM technology also offers migration of workloads from renewable energy–based CDCs to CDCs at another site utilizing waste heat. To balance the workload demand and renewable energy, VM-based workload migration and consolidation techniques provide virtual resources using few physical servers. To optimize virtualization performance, *storage* from one running server to another can be migrated without affecting VM workload execution. Waste heat utilization and renewable energy resource alternatives are harnessed by VM migration techniques to enable sustainable cloud computing. It is a great challenge for VM migration techniques to improve energy utilization and network delay while migrating workloads between resources distributed geographically. Increasing the size of the VM consumes more energy, which can increase service delay. To solve this problem, point-to-point communication is required for VM migration using a WAN.

Waste Heat Utilization

The *heat transfer* and *cooling mechanism* model plays an important role in effectively utilizing waste heat. Due to the consumption of large amounts of energy, CDCs act as a heat generator. Their vapor absorption–based cooling systems can use waste heat by utilizing heat while evaporating. Vapor absorption–based free cooling techniques can make the value of power usage efficiency (PUE) ideal by neutralizing cooling expenses. Low-temperature areas can use the heat generated by the CDC for heating purposes. Power densities of servers can be increased by using stacked and multi-core server designs, which further increase cooling costs. The energy efficiency of CDCs can be improved by reducing the energy usage in cooling. To reduce cooling costs, CDCs can be placed in areas with availability of free cooling resources.

Thermal-Aware Scheduling

The important components of thermal-aware scheduling are *architecture* and *scheduling mechanisms*. Architecture can be single-core or multi-core, while scheduling mechanisms can be reactive or proactive. Heating problems during the execution of workloads reduce the efficiency of cloud data centers. To solve

CDC heating problems, thermal-aware scheduling is designed to minimize the cooling setpoint temperature, hotspots, and thermal gradient. Thermal-aware scheduling is economical and effective as compared to heat modeling. The energy consumption of CDCs can be minimized by activating servers adjacent to each other in a rack or chassis, but power density increases, which creates heat concentration. To solve this problem, a cooling mechanism is required. Therefore, there is a need for effective thermal-aware scheduling techniques, which can execute workloads with minimum heat concentration and dissipation. This also reduces the load on the cooling mechanism, saving electricity. Complexity of scheduling and monitoring is increased due to the variation of temperatures of the servers in the CDC, which also causes vagueness in thermal profiling. To solve this problem, there is a need for dynamically updated thermal profiles instead of static ones; dynamic profiles update automatically and provide more accurate temperature values. Existing thermal-aware techniques focused on reducing PUE can be found, but a reduction in PUE may not reduce total cost of ownership.

Renewable Energy

Energy source (solar or wind), energy *storage device* (net-metering or batteries), and *location* (offsite or onsite) are important components of renewable energy, which can be optimized. The main challenges of renewable energy are unpredictability and capital cost of green resources. Workload migration and energy-aware load-balancing techniques address the issue of unpredictability in the supply of renewable energy. For the most part, commercial CDC sites are located away from abundant renewable energy resources. Consequently, CDCs need to be moveable so they can be placed closer to renewable energy sources to be cost-effective. Furthermore, carbon usage efficiency can be reduced by adding renewable energy resources. Adoption of renewable energy in CDCs has the challenge of high capital cost.

Resources Targeted in Energy Management

Many solutions have been proposed to improve the energy efficiency of CDCs. The energy consumption of the *processor*, *memory*, *storage*, *network*, and *cooling* of CDCs is reported as 45%, 15%, 10%, 10%, and 20%, respectively.[1] The processor consumes the most energy, followed by cooling. Power-regulation approaches increase energy consumption during workload execution, which affects the resource utilization of CDCs. DVFS solves the problem of resource utilization but the switching of resources between high-scaling and low-scaling modes increases response time and service delay, which can violate the SLA. Putting servers in sleeping mode or turning servers on/off affects the reliability of system components such as storage. Efforts to improve CDC energy efficiency affect the resource utilization, reliability, and server performance. The bin packing solution has been used in existing energy-aware resource management techniques to allocate resources for the execution of workloads. But resource allocation faces two problems: (1) underutilization of resources (i.e., resources are reserved in advance, but the resource requirement is

[1] See "Renewable and Cooling Aware Workload Management for Sustainable Data Centers" and "Sustainable and Renewable Energy: An Overview of the Application of Multiple Criteria Decision Making Techniques and Approaches."

lower than resource availability, increasing cost) and (2) overutilization of resources (i.e., a large number of workloads are waiting for execution due to unavailability of a sufficient amount of resources). Several methods have been proposed to control energy consumption by scaling down high-voltage supply, but the best way is to exploit the stall time.

Capacity Planning

Cloud service providers must maintain effective and organized capacity planning to attain a solid ROI. Capacity planning can be done for *power infrastructure*, *IT resources*, and *workload*. The SLA should define service quality parameters to ensure backup/recovery, storage, and availability that improves user satisfaction and attracts more customers in the future. There is a need to consider important utilization parameters per application to maximize the utilization of resources through virtualization by finding applications that can be merged. Merging of applications improves resource utilization and reduces capacity cost. For efficient capacity planning, cloud workloads should be analyzed before execution in order to finish execution for deadline-oriented workloads. To manage power infrastructure effectively, VM migration should be provided for the migration of workloads or machines to successfully complete the execution of workloads with minimum usage of resources, which improves the energy efficiency of CDCs. Thus, there is a need for effective capacity planning for data storage and processing effectively at lower cost.

Summary

In this *Update*, we identified the need for and the issues surrounding the sustainability of cloud computing environments. We proposed a conceptual model for the holistic management of resources to decrease the carbon footprints of cloud data centers, making cloud services more energy-efficient and sustainable. Holistic management improves the energy efficiency of the power infrastructure and cooling devices by integrating them into the energy-aware resource management technique applied with the equipment.

About the Authors

Rajkumar Buyya is a Redmond Barry Distinguished Professor and Director of the Cloud Computing and Distributed Systems Laboratory at the University of Melbourne, Australia. Dr. Buyya is an IEEE Fellow, earned recognition as a Web of Science Highly Cited Researcher (2016, 2017), and received the Scopus Researcher of the Year with Excellence in Innovative Research Award (2017) for his outstanding contributions to cloud computing. He can be reached at rbuyya@unimelb.edu.au.

Sukhpal Singh Gill is a Postdoctoral Research Fellow of the Cloud Computing and Distributed Systems Laboratory at the University of Melbourne, Australia. He can be reached at sukhpal.gill@unimelb.edu.au.