Guest Editors' Introduction: Special Issue on Green and Energy-Efficient Cloud Computing Part II

Ricardo Bianchini, *Fellow, IEEE*, Samee U. Khan, *Senior Member, IEEE*, and Carlo Mastroianni, *Member, IEEE*

CLOUD Computing has had a huge commercial impact and has attracted the interest of the research community. Public clouds allow their customers to outsource the management of physical resources, and rent a variable amount of resources in accordance to their specific needs. Private clouds allow companies to manage on-premises resources, exploiting the capabilities offered by the cloud technologies, such as using virtualization to improve resource utilization and cloud software for resource management automation. Hybrid clouds, where private infrastructures are integrated and complemented by external resources, are becoming a common scenario as well, for example to manage load peaks.

Cloud applications are hosted by data centers whose size ranges from tens to tens of thousands of servers, which raises significant challenges related to energy and cost management. It has been estimated that the Information and Communication Technology (ICT) industry alone is responsible for 2-3 percent of the global greenhouse gas emissions. Therefore, we must find innovative methods and tools to manage the energy efficiency and carbon footprint of data centers, so that they can operate and scale in a cost-effective and environmentally sustainable manner. These methods and tools are often categorized as Data Center Infrastructure Management (DCIM) to monitor, control, and optimize data centers with extensive automation. DCIM must also effectively manage the quality of service provided by the data center, since cloud customers require high reliability, availability, usability, and low response times.

While significant advancements have been made to increase the physical efficiency of power supplies and cooling components that improve the Power Usage Effectiveness (PUE) index, such improvements are often circumscribed to the huge data centers run by large cloud companies. Even stronger effort is needed to improve the data center computational efficiency, as servers are today highly underutilized, with typical operating range between 10 and 30 percent. In this respect, advancements are needed both to improve the energy-efficiency of servers and to dynamically consolidate the workload on fewer, and better utilized, servers.

This special issue has offered the scientific and industrial communities a forum to present new research, development, and deployment efforts in the field of green and energy-efficient Cloud Computing. Indeed, the special issue attracted a large number of good quality papers. After two or, in some cases, three rounds of reviews—each involving at least three expert reviewers—18 papers have been selected for publication among the 44 initially submitted. The accepted papers have been split into two issues of this journal. The first issue, published as Vol 4, No. 2, included nine papers that focus on the opportunities offered by the modern virtualization technology for reducing energy consumption and carbon emissions, through techniques and methods that aim to achieve optimal allocation and scheduling of virtual machines (VMs), both in single platforms and in geographically distributed scenarios involving multiple data centers. This issue includes nine papers that are more specifically devoted to the energy-efficient management of the physical infrastructure of data centers and cloud facilities, renewable energy, and networking and storage systems.

The first two papers of this issue explore the opportunities that network virtualization offers for performance improvement and energy efficiency in cloud data centers. In [1], Ghazisaeedi and Huang illustrate how network virtualization technology helps consolidate and optimize the infrastructure by allowing the coexistence of multiple virtual networks over a single physical network. The work presents a methodology and a set of algorithms that are able to reduce the power consumption in physical links during off-peak time. Specifically, the authors combine coarse-grained/centralized optimization, where the granularity is at the virtual link level, with fine-grained/local optimization, where the granularity is the traffic capacity that must be allocated to a virtual link. They formulate the problem as an Integer Linear Program (ILP), and devise a novel heuristic algorithm to increase scalability in large network sizes.

Yang and colleagues address virtual data center embedding in a virtual network environment [2]. The problem

- *R. Bianchini is with Microsoft Research, Redmond, WA , 98052.*
- *S.U. Khan is with the Department of Electrical and Computer Engineering, North Dakota State University, Fargo, ND 58108-6050.*
  E-mail: samee.khan@ndsu.edu.
- *C. Mastroianni is with the Institute for High Performance Computing and Networking, ICAR-CNR, via P Bucci 41C, Rende (CS) 87036, Italy.*
  E-mail: mastroianni@icar.cnr.it.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TCC.2016.2641618

involves mapping physical resources onto virtual resources with node and communication link constraints. In its general form, this problem is known to be NP-hard, even if the mapping is done offline. The authors address this issue by developing two closed-form solutions, which they verify using an elaborate benchmarking setup that captures the system dynamics and quantifies the network scale and topology changes. Their results indicate significant energy savings when mapping is done judicioulsly.

In [3], the focus in on air conditioners, which are one of the main energy consumers in data centers. The authors observe that private data centers are often placed within larger buildings, and therefore are adjacent to office space. They introduce a system that utilizes exhaust heat from servers to condition the humidity and air temperature in office space. As a result, energy consumption by air conditioners installed in the office space can be decreased. They also propose a tandem equipment arrangement that divides aisles into three classes: cold, hot, and super-hot. The high-temperature air in the super-hot aisle is collected by the exhaust heat reuse system. Numerical simulation shows that the total energy consumption in the proposed data center architecture is 27 percent lower than that of a conventional architecture.

Yu and Pan study energy-aware dynamic server provisioning in distributed caching systems based on consistent hashing [4]. The key challenge is devising workload consolidation algorithms that conserve energy while maintaining a high quality of service (high caching performance and stability). To address this challenge, the authors formulate a stochastic network optimization problem, and propose an epoch-based online algorithm that controls the workload consolidation and request dispatching based on the epochs' queue lengths. The evaluation involves extensive simulations of hundreds of servers and realistic access traces, and shows positive results.

In [5], the authors offer an interesting perspective on the need for rationing water consumption, especially when facing drought emergencies. The impact of draughts on data centers can be severe because an enormous amount of water is needed both for the production of electricity and for the correct operation of cooling systems. This work presents a software-based approach that aims to optimize the workload management and survive droughts by keeping the long-term overall water footprint under a cap. The approach exploits the inherent spatial and temporal diversities of data centers' water efficiencies and is based on two main techniques: geographic load balancing, i.e., dynamically dispatch workloads to data centers, and power proportionality, i.e., dynamically turn on/off servers in accordance with workloads. Analytical and simulation results show that the approach can cut water consumption by 20 percent while only incurring a negligible operational cost increase when compared to state-of-the-art cost-minimizing but water-oblivious solutions.

In [6], the authors discuss distributed storage systems and the related issues, such as performance and energy efficiency. They point out that the energy efficiency of a distributed storage system is highly dependent on several factors, such as data volume, data access frequency, and data redundancy. Consequently, the authors analyze a heterogeneous distributed storage system with data being categorized in a set number of classes and stored in accordance with such a categorization. The analysis reveals that the energy efficiency is closely related to the latency. Using erasure codes, the authors study a queuing model for the distributed storage system to derive the lower and upper bounds on the average latency for the various data classes. The study reveals some intricate relations between the energy efficiency and coding rate, service redundancy, and the number of redundant data requests.

Lee and colleagues highlight the importance of resource allocation mechanisms that are cognizant of the thermal signatures of data centers that run virtual high-performance computing systems in the cloud [7]. By mitigating thermal imbalance, hot spots can be avoided, which directly (by load balancing) and indirectly (by cooling) may reduce the energy consumption of the data center. Consequently, the authors propose a proactive thermal-aware resource management technique that aids in the minimization of hot spot creation. They benchmark the proposed methodology with high-performance computing workloads under single and federated data center scenarios.

In [8], the authors detail a new service level agreement setup that incorporates the feasibility of utilizing renewable energy resources. The proposed method (a) develops the concept of virtualized green energy to circumvent the uncertainty in the availability of renewable energy sources; (b) incorporates specific language in the service level agreement to translate the expectations (user and provider) in the presence of renewable energy sources; and (c) develops a resource allocation methodology that matches the cloud resources to the expectations laid out in the modified service level agreement. The authors test the proposed methodology using PlanetLab and SPECpower characterizations.

The goal of [9] is to develop an efficient strategy to reduce the energy consumption of application servers in cloud environments, with the minimum performance degradation. The strategy is based on the analysis of application behavior and on the observation that the processor frequency can be adjusted depending on the specific execution phase. The approach uses the information already available in the Java Virtual Machine to detect execution phases at run-time and exploit the application behavior. Experimental results show that the use of the proposed power-saving strategy leads to a significant reduction of the energy consumed by long running application servers, between 18 and 24 percent, without performance degradation.

We hope that this special issue will help the community understand the state of the art, determine future goals, and define architectures and technologies that will foster the adoption of greener and more efficient cloud resources. We would like to thank all the researchers who submitted papers, and all the reviewers who helped improve the quality of the published papers. Finally, we also warmly thank the TCC administrator, Ms. Joyce Arnold, and the Journals Coordinator, Ms. Erin Espriu, for supporting the special issue.

References

[1] E. Ghazisaeedi and C. Huang, "Off-peak energy optimization for links in virtualized network environment," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[2] Y. Yang, X. Chang, J. Liu, and L. Li, "Towards robust green virtual cloud data center provisioning," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[3] Y. Taniguchi, et al., "Tandem equipment arranged architecture with exhaust heat reuse system for software-defined data center infrastructure," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[4] B. Yu and J. Pan, "Optimize the server provisioning and request dispatching in distributed memory cache services," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[5] M. Islam, S. Ren, G. Quan, M. Shakir, and A. Vasilakos, "Water-constrained geographic load balancing in data centers," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[6] A. Kumar, R. Tandon, and T. Clancy, "On the latency and energy efficiency of distributed storage systems," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[7] E. Lee, H. Viswanathan, and D. Pompili, "Proactive thermal-aware resource management in virtualized HPC cloud datacenters," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[8] S. Hasan, Y. Kouki, T. Ledoux, and J. Pazat, "Exploiting renewable sources: When green SLA becomes a possible reality in cloud computing," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

[9] K.-Y. Chen, J. Chang, and T.-W. Hou, "An energy-efficient java virtual machine," *IEEE Trans. Cloud Comput.*, vol. 5, no. 2, 2016.

Ricardo Bianchini received the PhD degree in computer science from the University of Rochester in 1995. He was an associate professor of computer science with the Federal University of Rio de Janeiro until 1999, and a professor of computer science with Rutgers University until 2015. He is currently Microsoft's chief efficiency strategist. His main interests include cloud computing, and power/energy/thermal management of datacenters. In fact, he is a pioneer in datacenter energy management, energy-aware storage systems, energy-aware load distribution across datacenters, and leveraging renewable energy in datacenters. He has published eight award papers, and has received the CAREER award from the National Science Foundation. He is currently an ACM Distinguished Scientist and a Fellow of the IEEE.

Samee U. Khan received the BS degree from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, in 1999, and the PhD degree from the University of Texas, Arlington, Texas. Currently, he is an associate professor of electrical and computer engineering at the North Dakota State University, Fargo, North Dakota. His research interests include optimization, robustness, and security of: cloud, grid, cluster and big data computing, social networks, wired and wireless networks, power systems, smart grids, and optical networks. His work has appeared in more than 300 publications. He is on the editorial boards of leading journals, such as the *IEEE Access*, the *IEEE Cloud Computing*, the *IEEE Communications Surveys and Tutorials*, and the *IEEE IT Pro*. He is a fellow of the Institution of Engineering and Technology (IET, formerly IEE), and a fellow of the British Computer Society (BCS). He is an ACM Distinguished lecturer, a member of the ACM, and a senior member of the IEEE.

Carlo Mastroianni received the Laurea degree and the PhD degree in computer engineering from the University of Calabria, Italy, in 1995 and 1999, respectively. He is a researcher at the Institute of High Performance Computing and Networking of the Italian National Research Council, ICAR-CNR, in Cosenza, Italy, since 2002. Previously, he worked at the Computer Department of the Prime Minister Office, in Rome. He co-authored more than 100 papers published in international journals, among which IEEE/ACM TON, IEEE TCC, IEEE TEVC and ACM TAAS, and conference proceedings. He edited special issues for the *Journals Future Generation Computer Systems*, the *Journal of Network and Computer Applications*, the *Computer Networks Multiagent and Grid Systems*. His areas of interest are cloud computing, P2P, bio-inspired algorithms, multi-agent systems. He is a member of the IEEE.