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

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 present issue includes 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. A forthcoming issue will include nine papers that are more specifically devoted to the efficient management of the physical infrastructure of data centers and cloud facilities.

K. Li in [1] develops a queuing model for a multicore system with workload-dependent dynamic power management. The author derives for that model the necessary and sufficient conditions that allow the average task response time to be minimized. The most impressive result of this work is the establishment that the power consumption reduction, constrained by the performance guarantees, can be studied in a similar way as performance improvement (average task response time reduction) subject to power constraints. The work also reports several speed schemes to demonstrate the fact that for the same average power consumption, it is possible to design a multicore processor so that the average task response time is shorter than a multicore processor with uniform clock rate.

In [2], an eco-aware approach is presented that relies on the definition, monitoring and utilization of energy and CO_2 metrics combined with the use of innovative application scheduling and runtime adaptation techniques. The aim is to optimize energy consumption and CO_2 footprint of cloud applications as well as the underlying infrastructure. The authors have developed a layered approach for the

- *R. Bianchini is with Microsoft Research, Redmond, WA.*
- *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, 87036, Rende, (CS), 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.2015.2506298

definition of usage and power/energy metrics at three different layers, the application layer, the VM layer and the infrastructure layer. Starting from the real-time analysis of those metrics, they have defined an eco-aware scheduler that drives the allocation and execution of VMs within a single site as well as in federated cloud infrastructure, and is capable of obtaining reduction of carbon emissions up to 80 percent.

While there is a vast literature on the efficient management of cloud platform at the virtualization level, there is still much room for improving the environmental impact through a better management at the application level. For example, the correlation between the characteristics of applications and the behavior of the VMs that host the applications should be better investigated and exploited. This is the intent of the study presented in [3]: the authors propose an approach to control the applications during their life-cycle, from design to execution. At design time, the approach suggests the optimal set of VMs to be deployed according to the application profile; at run time, a set of adaptation strategies are enacted to reduce CO_2 emissions. These strategies are based on the information available at the application level to reduce the environmental impact once the applications are deployed in a virtualized infrastructure. The potential for improvement, when using the presented approach, can lead to a 60 percent reduction of CO_2 emissions, without degrading performance.

Tracking energy consumption and CO_2 footprint of cloud applications becomes even more difficult when applications are complex and span multiple cloud computing platforms. The authors of [4] focus on the energy-efficient execution of scientific workflows that need to be deployed across multiple data centers due to their large-scale characteristics. The optimal allocation of virtual machines must consider that workflow tasks have dependencies and communication constraints which make them differ significantly from unrelated tasks. The solution presented in the paper, *EnReal*, addresses such challenges by leveraging the dynamic deployment of virtual machines: an energy consumption model is devised, and a corresponding energy-aware resource allocation algorithm is proposed for virtual machine scheduling. The approach has been evaluated through a wide set of simulation experiments executed over the CloudSim framework.

The energy-efficient management of geographically distributed data centers is also the subject of [5]. The authors focus on the significant impact of "geotemporal inputs", i.e., the time- and location-dependent factors that may impact energy consumption. Among such factors, they consider real-time electricity pricing enabled by the deregulated electricity market, the cooling efforts needed at different sites and different times, and the availability of renewable energy. The difficult problem of scheduling the VMs in a geographical context is tackled through a two-stage approach, which combines best-effort global optimization, driven by genetic algorithms, with deterministic local optimization for constraint satisfaction. A reported simulation study based on real traces of temperatures and electricity prices shows that considerable energy cost savings, of up to 28.6 percent, are achievable compared to a baseline control method that applies VM consolidation without considering geotemporal inputs.

In [6], the classical problem of VM consolidation, i.e., reduce the energy consumption of a data center by packing the running VM instances to as few physical machines as possible, is considered. Since VM consolidation is an applied form of the NP-hard bin packing problem, it cannot be solved in large data centers using classical centralized approaches. In [6], the problem is tackled using an original approach, partly inspired by the state-of-the-art in the peer-to-peer field. Specifically, the physical machines of the data center self-organize in a hypercube overlay network, and each host autonomously decides if some VMs should be offloaded to improve the consolidation ratio. In this case, hypercube connections are used to choose the target hosts and migrate the VMs. The benefits of the approach, in terms of energy-efficiency and scalability, derive from the properties of the overlay structure, i.e., low number of connections, reduced number of exchanged messages and strong resilience to high churn rates.

Dai et al. also investigate the problem of energy-aware placement of VMs onto a subset of (active) servers in [7]. They formulate the problem using integer programming, prove that it is NP-hard, and propose two greedy approximation algorithms to reduce the energy consumption while satisfying the tenants' service level agreements (SLAs). Their results demonstrate that the algorithms efficiently produce placements that are close to optimal in terms of energy consumption, while satisfying all SLAs.

In a similar context, Goudarzi and Pedram propose a scalable solution to the VM placement problem in [8]. Specifically, they propose a hierarchical placement framework that accounts for both server and cooling power in seeking lower operational costs. The results show that their hierarchical resource manager computes VM placements significantly faster than centralized solutions, while also lowering operational costs.

In [9], Luo et al. propose eCope, a general framework for adapting the power state of hardware components, according to the system's workload characteristics. They demonstrate the use of eCope for three systems: the Taobao distributed file system, the MySQL database, and the Apache Web server. Their results show significant dynamic power savings at the cost of only a slight 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] K. Li, "Improving multicore server performance and reducing energy consumption by workload dependent dynamic power management," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 122–137, 2016.

[2] U. Wajid, C. Cappiello, P. Plebani, B. Pernici, N. Mehandjiev, M. Vitali, M. Gienger, K. Kavoussanakis, D. Margery, D. Perez, and P. Sampaio, "On achieving energy efficiency and reducing CO2 footprint in cloud computing," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 138–151, 2016.

[3] C. Cappiello, N. Ho, B. Pernici, P. Plebani, and M. Vitali, "CO2-aware adaptation strategies for cloud applications," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 152–165, 2016.

[4] X. Xu, W. Dou, X. Zhang, and J. Chen, "EnReal: An energy-aware resource allocation method for scientific workflow executions in cloud environment," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 166–179, 2016.
[5] D. Lucanin and I. Brandic, "Pervasive cloud controller for geotemporal inputs," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 180–195, 2016.
[6] M. Pantazoglou, G. Tzortzakis, and A. Delis, "Decentralized and energy-efficient workload management in enterprise clouds," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 196–209, 2016.
[7] X. Dai, M. Wang, and B. Benasou, "Energy-efficient virtual machines scheduling in multi-tenant data centers," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 210–221, 2016.
[8] H. Goudarzi and M. Pedram, "Hierarchical SLA-driven resource management for peak power-aware and energy-efficient operation of a cloud datacenter," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 222–236, 2016.
[9] B. Luo, S. Wang, W. Shi, and Y. He, "eCope: Workload-aware elastic customization for power efficiency of high-end servers," *IEEE Trans. Cloud Comput.*, vol. 4, no. 2, pp. 237–249, 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 at the Federal University of Rio de Janeiro until 1999, and a professor of computer science at 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 in 1999 from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan and the PhD degree in 2007 from the University of Texas, Arlington, TX. Currently, he is an associate professor of electrical and computer engineering at the North Dakota State University, Fargo, ND. 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 *IEEE Access*, *IEEE Cloud Computing*, *IEEE Communications Surveys and Tutorials*, and *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 coauthored more than 100 papers published in international journals, among which *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Evolutionary Computation*, and *ACM Transactions on Autonomous and Adaptive Systems*, and conference proceedings. He edited special issues for the journals *Future Generation Computer Systems*, *Journal of Network and Computer Applications*, *Computer Networks Multiagent and Grid Systems*. His areas of interest are cloud computing, P2P, bio-inspired algorithms, and multi-agent systems. He is a member of the IEEE.