Multilevel resource allocation for performance-aware energy-efficient cloud data centers

Fábio Diniz Rossi*, Paulo Silas Severo de Souza†, Wagner dos Santos Marques†,
Marcelo da Silva Conterato†, Tiago Coelho Ferreto†, Arthur Francisco Lorenzon‡, Marcelo Caggiani Luizelli‡
*Federal Institute of Education, Science, and Technology Farroupilha (IFFar) - Alegrete - Brazil
†Pontifical Catholic University of Rio Grande do Sul (PUCRS) - Porto Alegre - Brazil
‡Federal University of Pampa (UNIPAMPA) - Alegrete - Brazil
Email: fabio.rossi@iffarroupilha.edu.br

***Abstract*—The massive power consumption of data centers has been a recurring concern in current research. In cloud environments, lots of methods are being adopted that aim for energy efficiency. However, although such methods enable the decrease in power consumption, they regularly affect application performance. In this paper, we present a multilevel resource allocation approach towards dynamic network bandwidth at the physical substrate, managing different power-saving states and workload allocation at the cloud infrastructure at the same time employ virtual machine allocation and selection policies at the cloud platform. In order to evaluate our approach, tests were carried out in a simulated environment using scale-out application on a dynamic cloud infrastructure. Results showed that our proposal presents a better balance regarding a more energy-efficient data center with a smaller impact on application performance when compared with other works discussed in the literature.**

Keywords*—*Cloud Application Performance; Data Center; Energy-Efficient Management.

I. Introduction

Cloud computing allows access to data, computation, and applications as services from anyplace by the Internet. Customers are not tied to a physical infrastructure because their data and applications are processing by services. Also, cloud customers only pay for that they use (pay-per-use model), without overspending and acquiring unnecessary resources. Furthermore, requirements such as reliability, security, availability, fault tolerance, scalability, and sustainability boost cloud computing as a de facto standard in the industry [1]. Nevertheless, due to the movement of applications from traditional models to cloud environments, data centers started to increase the number of available resources to meet this increased demand. Although the adjustment in the volume of data center resources is a fundamental process, the higher utilization of resources requires a more significant amount of energy to keep them active, impacting on sustainable issues [2]. Power consumption boosts the heat dissipation by computing equipment, which raises the emission of gases that cause the greenhouse effect. Therefore, with the increment of the devices number in data centers in addition to the cooling needs, power consumption has become an environmental and economic issue. Besides, the increased processing power is only possible by the increase in energy consumption. In this sense, it looks to be reasonable that energy savings necessarily imply in decreased performance of cloud servers. Although, the impact of energy savings in cloud data centers is one of the most studied topics today [3]. Enhanced management of resources presents the potential to decrease energy consumption, and then decrease the impact on the atmosphere. In this way, several energy-saving methods have been proposed [4]. Meanwhile, applications performance metrics became a competing circumstance under the cloud service providers' perspective, and on the customer view, such metrics impact the quality of experience (QoE).

This work proposes to improve the trade-off between power savings and applications performance, operating both at the cloud infrastructure layer and at the cloud platform layer. At the infrastructure layer, we tune the hosts over different sleep states, managing them between states in a way that does not lose performance. At the platform layer, we allocate virtual machines according to different types of policies. Below these levels, we still propose dynamic communication channels where virtual machines can travel faster from a host to another. This multilevel resource management allows more energy to be saved, with less impact on applications performance. We implemented the proposed approach in a simulated cloud environment and evaluated the provided improvement in terms of energy savings and application performance. The results showed that our approach can save energy up to 33% when compared to the power-agnostic approach, with the advantage of losing less performance when compared to previous work. The direct benefit driven by this work is the promotion of the equilibrium between energy savings and applications performance in cloud environments. Beyond this important contribution, an anticipated impact of this work considers indirect environmental benefits. The reminder of this paper is structured as follows. The problem statement is discussed in Section 2. Section 3 drives the motivation for this work by uncovering open problems and discussing value potential. We review and put our work in the context of related work in Section 4. We present a quantitative evaluation of the benefit of our approach simulated workloads in Section 5 and conclude with summary and elevation of the findings in Section 6.

II. Problem Statement

In the computational area, data centers are the most energy-consuming infrastructures generated by traditional sources from fossil fuels. Such a practice has a direct impact on environmental issues such as global warming and the greenhouse effect. International agreements reinforced by exemption from taxes have guided the reduction of energy consumption in these facilities. One of the most feasible ways to reduce the energy consumption is to adopt strategies to allocate application on the resources in an intelligent way. The problem is the overhead such strategies impose on others equally essential metrics in an enterprise environment. One of the most critically affected metrics is the application's performance which directly involves customers' quality of experience. While cloud services

978-1-7281-2999-0/19/$31.00 ©2019 IEEE

allow access by the Internet, application performance becomes a decisive factor in customer loyalty to the service offered. As an example of power-performance trade-off analysis, let us consider the performance (e.g., execution time) and power consumption of a cloud environment. The higher the application's performance, the more resources are used and the higher the power consumption. The higher the power consumption, the more significant amount of resources is in use, providing higher application's performance. Besides, several works propose energy savings through the use of different standby states, such as standby, hibernate, or even shut down idle hosts. Another limiting factor that is not taken into account by most works is the time elapsed between state transitions.

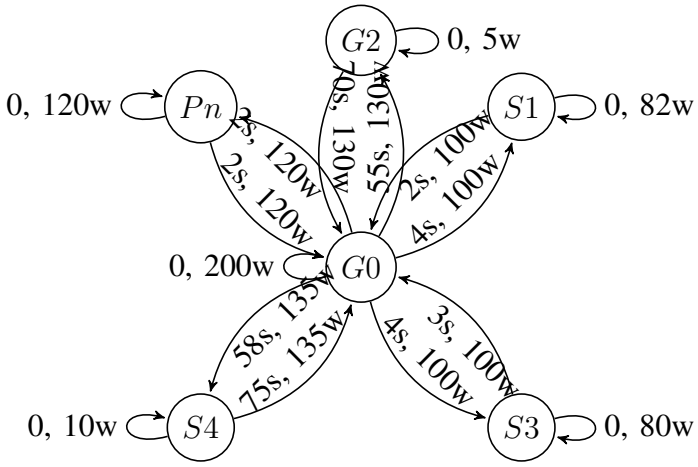

Fig. 1: Evaluation of transition time and energy consumption during the ACPI transitions. G0: the system and user threads are running (working), G2: the system consumes a minimal amount of power, user mode threads and system processes are not running, and the system context is not saved (Soft Off), S1: no system context is lost, S3: CPU, system cache, and chip set context are lost, S4: powered off all devices, Pn: the processor performance capability is at its minimum level and consumes minimal power while remaining in an active state.

To investigate new states that can be used to replace the idle state, we conduct evaluations of energy consumption in each sleep state. The experiments were conducted in a host with Intel Xeon processors E5645 2.40 GHz, 12/24 cores, 64 MB L3 Cache - Ubuntu Linux 12.04 LTS - Kernel version 3.13.0. Since the focus of our work is to improve the trade-off between energy savings and applications performance, we also evaluated the time needed to complete each state transition. Results of such tests are depicted in Figure 1. Our proposal uses such a model to implement the cloud orchestration. Therefore, the approach we propose can be understanding as solving a bi-objective optimization problem, where we attempt to minimize two objectives, namely the application performance and the cloud environment power consumption. A scatter plot of the objective values corresponding to configurations can give cloud providers an overview of how power and performance interact in the cloud environment. It can support to design optimization algorithms for an efficient energy-aware approach that can handle a broad assortment of power-performance requirements.

III. Related Work

In order to address the Pareto problem, modern virtualized environments bring opportunities for power management through energy-aware strategies on idle hosts [5], [6], [7]. In [5], Jeffrey et al. propose an architecture for resource management in a hosting center operating system, which dynamically resizes the active server set to improve the energy efficiency of server clusters. Heath et al. [6] present the design of a cooperative Web server to optimize the request on heterogeneous cluster regarding many metrics, such as power, energy, throughput, and latency. In the same way, but considering parallel applications running on a heterogeneous cluster, Zong et al. [7] propose a scheduling algorithm that dynamically allocates parallel tasks on heterogeneous nodes in order to reduce the power consumption of the entire system. The question of minimization of operational costs through the decrease of power consumption in cloud environments is widely discussed in several works, such as Gao et al. [8]. In this work, a task graph workload model is used together with a energy cost minimization framework for the cloud services providers in order to maximize the energy efficiency while ensuring that user deadlines defined in service level agreements are met. Furthermore, Isci et al. [9] show that there is an opportunity for energy-savings strategies in these environments using the concept of sleep states, by exploiting low-power and low-latency power states in enterprise servers.

Min et al. [10] present a framework called energy efficient sleep-state selection. It dynamically decides the sleep state that minimizes the power consumption of the entire system based on standardized workloads for smartphones. Niyato et al. [11] proposed an energy management approach to adjust the number of active servers to provide energy savings while the performance requirements are met. Alvarruiz et al. [12] proposed a management method for clusters and clouds that saves power by turning off idle hosts via the network. Results showed energy-savings of 38% for cluster and 16% for the cloud, respectively. In the same way, Lefèvre and Orgerie [13] present a cloud architecture that saves power based on heuristics to turning on/off hosts and virtual machines migration. Results showed energy-savings of up to 25% when compared to an energy-agnostic environment. However, none of these works are concerned with the performance of applications. Besides, all cited works only act on one layer of the cloud environment, usually at the infrastructure layer. Beloglazov and Buyya [14] developed heuristics for virtual machines allocation in the cloud to saving power. Results showed power savings of up to 83% when compared to energy-agnostic scenarios, although they presented an insignificant impact on application performance. Zhu et al. [15] proposed dividing a cloud into four areas: busy, active idle, sleep, and shutdown. Results show such organization can decrease the power consumption in up to 84%, but with an impact on the execution time of up to 8.85%. Our proposal uses several aspects of the works cited, but also presents differences that will raise our performance and energy-savings results. First, we used the software-defined network (SDN) to manage network intermediary devices to extend channels through link aggregation to provide a channel with optimal bandwidth for virtual machine traffic between hosts. Also, our approach acts on the infrastructure layer, controlling the state transition of hosts, ensuring that a significant number of hosts can be turned off, but still maintaining a sufficient number of hosts to meet sudden resource demands. It performs improvements in the trade-off between energy savings and performance of cloud applications. On top of that, we have added virtual machine allocation policies over the active hosts, making better use of the available resources managed by the platform, which will boost energy savings, while maintaining the same small impact on application performance.

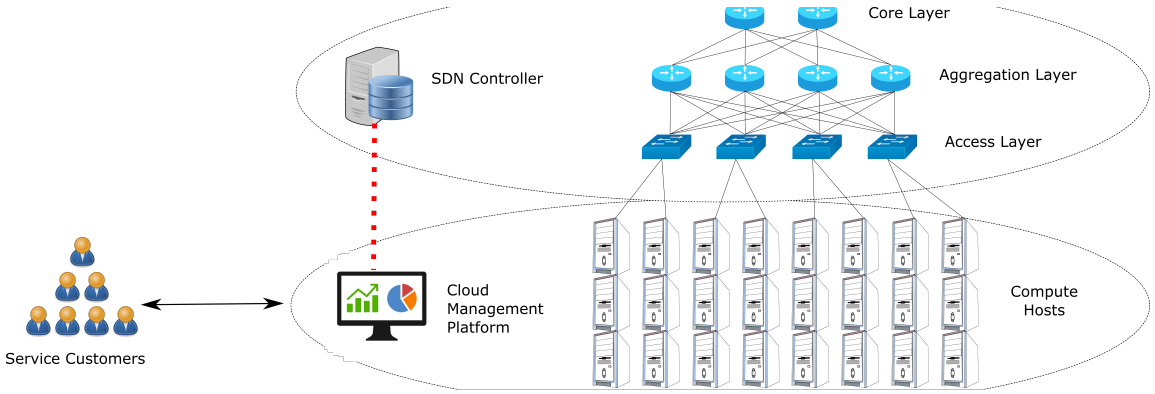

Fig. 2: Multilevel resource allocation approach on top of a three-tier network topology.

IV. Multilevel Resource Allocation Approach

Our approach combines a dynamic management of network bandwidth, a resource allocation strategy at the infrastructure layer, and virtual machine allocation policies on the resources at the platform layer.

A. Background

On the physical substrate of the data center, we use NeaReSt [16] (Network Rescaling Strategy) to dynamically manage network bandwidth when migrating virtual machines and data between hosts. The main goal of NeaReSt—a Network Rescaling Strategy—is improving the application's performance by the control of network bandwidth. NeaReSt targets cloud platform systems that may process a massive volume of data. Consequently, NeaReSt must have access to the physical substrate layer. Therefore, SDN is used to access, manage, and balance network use based on the application behavior. NeaReSt obtains information provided by the platform via SDN and then switches the network infrastructure settings, such as increasing or reducing network bandwidth via link aggregation. An SDN controller is required so that NeaReSt can have a holistic view of all network devices and network flows between such devices. SDN consists of technology that meets these requirements and enables flexible programmability. In this way, the network is being recognized as an elastic resource that can provide performance for applications by aggregating more than one physical channel, creating a virtual circuit based on the sum of the individual capacities of each physical channel. We used an SDN controller that implements the capability to view the traffic needs of hosts and determine the best alternative for data flow, based on predefined rules. NeaReSt is deployed in the SDN controller, and in addition to notifying to the switch the route that is to be used, it lets the combination of several physical links as one virtual channel from the origin to the destination to increase the network bandwidth and raise the data traffic. When the links are aggregated, and the switch tables understand the routing, the entire flow between source and destination is directed by this channel. When SDN controller realizes that the data flow is no longer occupying the virtual channel, such circuit is unbundled by releasing resources.

At the infrastructure layer we used e-eco [17] (energy-efficient cloud orchestrator) implementation. It consists in dividing the data center into sites with different energy consumption in order to save energy based on the idleness of some hosts caused by the floating behavior of applications over resources. Therefore, e-eco controls Advanced Configuration and Power Interface (ACPI) states [18], with the intention of separating hosts from a data center into three sites: (i) hosts over which the cloud applications are running; (ii) most idle hosts kept off, and (iii) a smaller set of hosts kept in an intermediate state of power consumption. It causes a growing demand to be met quickly by the intermediate site hosts, and the number of hosts maintained in this state is calculated through a mathematical model based on the periodicity of the increase or decrease in demand imposed by the costumers' requests. In order to move hosts between states, e-eco uses the Advanced Configuration and Power Interface (ACPI). Besides, e-eco controls migration of virtual machines and the use of Dynamic Voltage and Frequency Scaling (DVFS) over underutilized hosts. The possible e-eco transitions are hosts with running applications (G2), turned off hosts (G0), and an intermediate state for rapid deployment as well as less consumed energy (S3). The intermediate state is important to maintain Quality of Service when customers' requisitions increase rapidly. At the platform layer, we use in conjunction with NeaReSt and e-eco, and allocation and selection policies for virtual machines over the resources. The allocation policy decides how virtual machines will be organized over available resources. The selection policy decides which virtual machine could be reallocated, such as server consolidation.

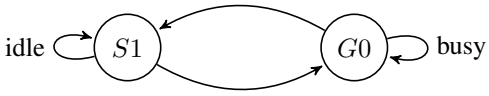

Fig. 3: Power-Agnostic Strategy employed on one host.

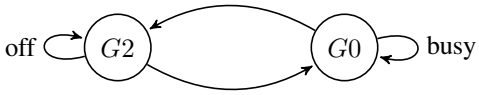

Fig. 4: Alvarruiz et al. strategy employed on one host.

B. Implementation

Figure 2 presents this work proposal architecture. Customer requests for the cloud service are answered by a cloud man-

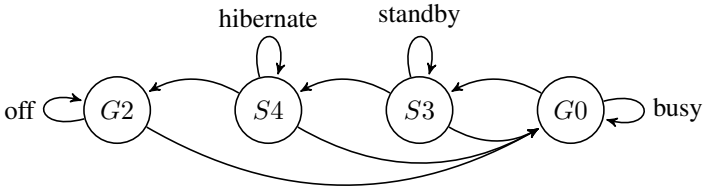

Fig. 5: Timeout strategy employed on one host.

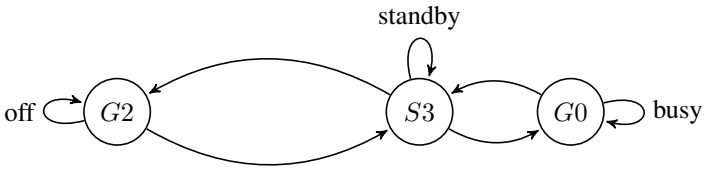

Fig. 6: e-eco strategy employed on one host.

agement platform (i.e., OpenStack), which balances the data load by replicating service over virtual machines. e-eco works along with cloud management platform, and separates compute hosts into three sites: (i) working, (ii) standby hosts expecting increased demand for consumer requests, and (iii) off. In order to improve resource utilization and consequently reduce power consumption, virtual machines that are instantiated on top of working hosts can be migrated between them. A common practice is to consolidate servers, wherein moments of low usage of resources, large amounts of virtual machines can be aggregated over a few hosts. Before virtual machine migrations occur, the cloud management platform sends a network message to the SDN controller with NeaReSt, notifying which host to migrate the source and destination host. Then, NeaReSt reconfigures the network so that there is an aggregation of links between the source and destination hosts, in order to optimize a communication channel so that virtual machine migration occurs more quickly. At times when virtual machines migrate via the network, aggregation of the physical channels into a single logical channel with the sum of all data transmission capabilities reduce the time required to send the data over the network, increasing the total cloud service performance. When the virtual machine migration is completed, NeaReSt releases the aggregated links, and waits for new messages coming from the cloud management platform. After, e-eco can release idle hosts and switches their states to standby or off. The implementation of the proposed architecture was performed on top of the CloudSim simulator [19] and uses two modules: the SDN module developed by [20], and the ACPI module developed by [18].

V. Evaluation and Discussion

The scale of a cloud data centers is not efficiently represented in academic environments [21]. For example, it is estimated that Amazon EC2 operates more than 450,000 servers [22]. We believe that the performance gains and power savings provided by this work remain for private cloud environments with a more significant number of hosts. Cameron [23] and Zomaya [24] refer to the increase of energy consumption concerning the scale of hosts in a near-linear or linear rate. Performance may also be supported by the scalability of the cloud environment [25]. It means that the energy savings and performance obtained are expected to be proportional, for each cloud usage rate, to a higher number of hosts.

A. Testbed

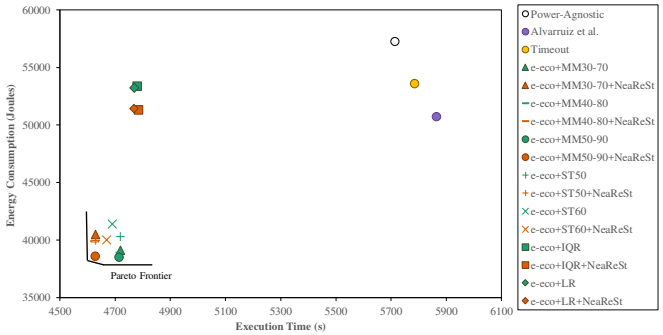

Fig. 7: Pareto frontier with respect to the performance and energy consumption of each approach

For the tests, we used the CloudSim Simulator [19]. For the simulated tests, we used a trace based on Beloglazov and Buyya [26]. We have simulated a data center three-tier network containing 1000 heterogeneous physical hosts. Each host is modeled to have one CPU core with the performance equal to 1000, 2000 or 3000 MIPS, 8 GB of RAM and 1 TB of storage. Each host consumes from 120W with 0% CPU utilization, up to 200W with 100% CPU utilization. Each virtual machine uses one CPU core with 250, 500, 750 or 1000 MIPS, 128 MB of RAM and 1 GB of storage. The customers submit requests for provisioning of 290 heterogeneous virtual machines that load the full capacity of the simulated environment. Each virtual machine runs a web-application or any kind of application with a variable workload, which is modeled to generate the utilization of CPU according to a uniformly distributed random variable. The application runs for 150,000 MI that is equivalent to 10min of the execution on 250 MIPS CPU with 100% utilization. Originally, the virtual machines are allocated according to the requested features assuming 100% CPU utilization. Each experiment has been run 10 times. It enables to evaluate our proposal in a larger environment and including cloud orchestration strategies and a more considerable amount of allocation policies. The orchestration strategies used are presented below: **Power-agnostic:** it is an environment that has no concern for energy savings. Figure 3 shows such behavior. **Alvarruiz et al.** [12]: it consists of an environment where hosts are kept in one of two states: running or off. In this scenario, VM consolidation and processors' frequency reduction are also applied to the hosts. Figure 4 shows such behavior. **Timeout** [27]: when the host in the G0 state becomes idle, it enters into the S3 state; the host returns immediately to the G0 state if it is requested; the host enters successively to a lower-power state if the timeout expires (300 seconds). Figure 5 shows such behavior. **e-eco** [17]: it divides the number of hosts into three states (running, standby, off), and the hosts are placed within these states depending on the application's demand, aiming for more significant energy savings with the least impact on performance. Figure 6 shows such behavior. A set of proposed policies tested along with orchestration strategies are Median Absolute Deviation and Minimum Migration Time (MM), Static Threshold and Random Selection (ST), Inter Quartile Range and Maximum Correlation (IQR), and Local Regression and Minimum Utilization (LR) [26]. Below is a description of

TABLE I: Evaluating Scenarios.

Strategy	Execution Time (s)	Energy Consumption (J)	EDP	SLA Violation (%)
Power-Agnostic	5715	57248	327172320	6
Alvarruiz et al.	5864	50720	297422080	39
Timeout	5785	53609	310128065	30
e-eco+MM30-70	4720	39114	184618080	5
e-eco+MM30-70+NeaReSt	4630	40500	187515000	4
e-eco+MM40-80	4715	38900	183417200	8
e-eco+MM40-80+NeaReSt	4628	39900	184657200	7
e-eco+MM50-90	4715	38500	181527500	13
e-eco+MM50-90+NeaReSt	4628	38600	178640800	10
e-eco+ST50	4720	40300	190216000	7
e-eco+ST50+NeaReSt	4630	39990	185153700	6
e-eco+ST60	4690	41400	194166000	8
e-eco+ST60+NeaReSt	4670	40040	186986800	7
e-eco+IQR	4780	53400	255252000	10
e-eco+IQR+NeaReSt	4785	51300	245470500	9
e-eco+LR	4770	53250	254002500	12
e-eco+LR+NeaReSt	4769	51450	245365050	10

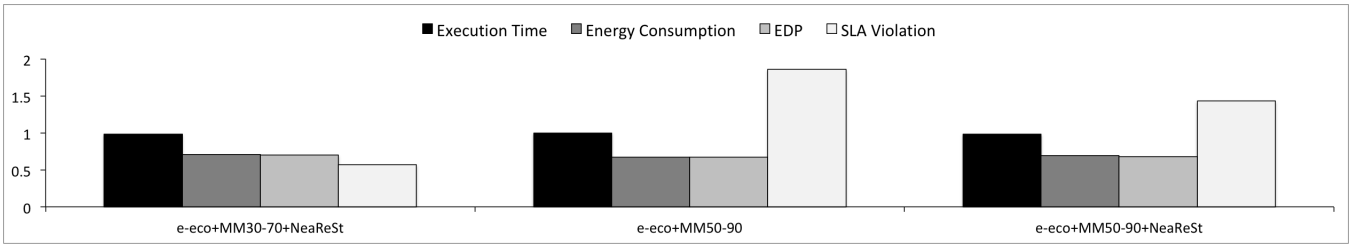

Fig. 8: More significant results normalized based on the power-agnostic strategy.

each of these policies: **MM:** it selects the minimum number of VMs needed to migrate from a host to lower the CPU utilization below the upper utilization threshold if the upper threshold is violated. We use different lower and upper limits for virtual machine migration, which are 30-70, 40-80, 50-90. **ST:** it relies on a random selection of a number of VMs needed to decrease the CPU utilization by a host below the upper utilization threshold. We use different upper limits for virtual machine migration, which are 50 and 60. **IQR:** it is a method for finding a dynamic threshold used to migrate virtual machines based on an estimate of variability, based on dividing a data set into quartiles. It is the difference between the upper and lower quartile in a data set. **LR:** it approximates the shorttime future processor utilization based on the history of usage in each host. It is employed in the live migration process to predict over-loaded and under-loaded hosts. A mix between e-eco, NeaReSt, and the virtual machine allocation policies presented compose the set of tests performed to determine which set can improve the trade-off between energy savings and performance of cloud applications in a data center.

B. Results

This section presents and discusses the results from the testbed described above, in terms of execution time, energy consumption, Energy-Delay Product (EDP), and Service Level Agreement (SLA) violation. Figure 7 presents the Pareto frontier regarding the performance (x-axis) and energy consumption (y-axis) of each method, while Table I depicts the raw numbers. As can be observed in Figure **??** and Table I, results show an execution time improvement in all e-eco plus NeaReSt plus Median Absolute Deviation and Minimum Migration Time (MM) virtual machine allocation policy sets when compared to power-agnostic strategy and other tested strategies and policies. It occurs because, in addition to e-eco maintaining an intermediary site for hosts that can respond to new demands quickly, migration of virtual machines is optimized by NeaReSt. The work of Rossi et al. [17] showed no improvement regarding performance because e-eco only showed energy savings gains without impacting on performance issues. Already this new perspective shows an increase in the performance of cloud applications by up to 19% when compared to the power-agnostic cloud environment. The most significant reductions in energy consumption also occurred in the set that used e-echo together with MM allocation policies and with/without NeaReSt. The best result was the one with MM policy using upper and lower limits of 50-90. However, this was also the test set that presented an SLA break above the others (13%). This particular case does not use NeaReSt, and the SLA break occurred because of the slower traffic of the virtual machines. However, this set is also the one that presents the most significant balance between energy savings and performance among all scenarios tested. When compared to power-agnostic, this scenario shows an improvement in EDP of 44%. The smallest SLA break occurred with the set using e-eco, NeaReSt, and MM using 30-70. This scenario reached a 33% better value than power-agnostic, and 69% better than the scenario with better EDP. Figure 8 summarizes the best scenarios, where we can see a very close value regarding execution time, power consumption, and consequently EDP. The main difference consists of the SLA breakdown, the factor that makes the best strategy among all tested ones is that it uses e-eco to manipulate the hosts in different states of

suspension, obeying limits to allow the migration of virtual machines between 30-70, and when such migration is required, there is an aggregation of links between hosts of origin and destination of such migration.

VI. Conclusion and future work

The benefits led by cloud computing has been promoting the establishment of data centers that support several different applications. Amongst the cloud benefits, smart management of resources is a crucial factor, as, through server virtualization, services can be scaled as they demand. Resource management as mentioned above impacts on operating costs for the service provider and one of the most significant is power consumption. Besides the consolidation of virtual machines intrinsically enabled by virtualized environments, several energy-saving techniques are used on cloud environments. This paper presents a cloud manager that improves the trade-off between energy savings and application performance through smart management of a set of power-saving methods. Results of our evaluation demonstrated that our approach could reduce energy consumption and SLA violations in up to 33% compared to power-agnostic approaches. Such a result showed that our proposal improves the trade-off between power savings and applications performance to enable a cloud environment that is at the same time economical and responsive. As future work, we intend to implement the solution on an OpenStack platform on a real cloud environment.

References

[1] R. Buyya, J. Broberg, and A. M. Goscinski, *Cloud Computing Principles and Paradigms*. Wiley Publishing, 2011.

[2] NRDC, "Data center efficiency assessment," Aug. 2014, http://www.nrdc.org/energy/files/data-center-efficiency-assessment-IP.pdf.

[3] F. D. Rossi, M. Conterato, T. C. Ferreto, and C. A. F. De Rose, "Evaluating the trade-off between dvfs energy-savings and virtual networks performance," in *Proceedings of the Thirteenth International Conference on Networks*, ser. ICN '14. Nice, France: IARIA, 2014, pp. 01–06.

[4] A. Grover, "Modern system power management," *Queue*, vol. 1, no. 7, pp. 66–72, Oct. 2003.

[5] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," in *Proceedings of the eighteenth ACM symposium on Operating systems principles*, ser. SOSP '01. New York, NY, USA: ACM, 2001, pp. 103–116.

[6] T. Heath, B. Diniz, E. V. Carrera, W. Meira, Jr., and R. Bianchini, "Energy conservation in heterogeneous server clusters," in *Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming*, ser. PPoPP '05. New York, NY, USA: ACM, 2005, pp. 186–195.

[7] Z. Zong, X. Qin, X. Ruan, K. Bellam, M. Nijim, and M. Alghamdi, "Energy-efficient scheduling for parallel applications running on heterogeneous clusters," in *Proceedings of the 2007 International Conference on Parallel Processing*, ser. ICPP '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 91–113.

[8] Y. Gao, Y. Wang, S. K. Gupta, and M. Pedram, "An energy and deadline aware resource provisioning, scheduling and optimization framework for cloud systems," in *Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, ser. CODES+ISSS '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 31:1–31:10.

[9] C. Isci, S. McIntosh, J. Kephart, R. Das, J. Hanson, S. Piper, R. Wolford, T. Brey, R. Kantner, A. Ng, J. Norris, A. Traore, and M. Frissora, "Agile, efficient virtualization power management with low-latency server power states," *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3, pp. 96–107, Jun. 2013.

[10] A. W. Min, R. Wang, J. Tsai, M. A. Ergin, and T.-Y. C. Tai, "Improving energy efficiency for mobile platforms by exploiting low-power sleep states," in *Proceedings of the 9th Conference on Computing Frontiers*, ser. CF '12. New York, NY, USA: ACM, 2012, pp. 133–142.

[11] D. Niyato, S. Chaisiri, and L. B. Sung, "Optimal power management for server farm to support green computing," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, ser. CCGRID '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 84–91.

[12] F. Alvarruiz, C. de Alfonso, M. Caballer, and V. Hernandez, "An energy manager for high performance computer clusters," in *IEEE 10th International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, july 2012, pp. 231 –238.

[13] L. Lefèvre and A.-C. Orgerie, "Designing and evaluating an energy efficient cloud," *Journal of Supercomputing*, vol. 51, no. 3, pp. 352–373, Mar. 2010.

[14] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, ser. CCGRID '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 826–831.

[15] H. Zhu, Y. Liu, K. Lu, and X. Wang, "Self-adaptive management of the sleep depths of idle nodes in large scale systems to balance between energy consumption and response times," in *Proceedings of the 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, ser. CLOUDCOM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 633–639.

[16] F. D. Rossi, G. D. C. Rodrigues, R. N. Calheiros, and M. D. S. Conterato, "Dynamic network bandwidth resizing for big data applications," in *2017 IEEE 13th International Conference on e-Science (e-Science)*, Oct 2017, pp. 423–431.

[17] F. D. Rossi, M. G. Xavier, C. A. D. Rose, R. N. Calheiros, and R. Buyya, "E-eco: Performance-aware energy-efficient cloud data center orchestration," *Journal of Network and Computer Applications*, vol. 78, pp. 83 – 96, 2017.

[18] M. G. Xavier, F. D. Rossi, C. A. F. De Rose, R. N. Calheiros, and D. G. Gomes, "Modeling and simulation of global and sleep states in acpi-compliant energy-efficient cloud environments," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 4, p. e3839, e3839 cpe.3839.

[19] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, no. 1, pp. 23–50, Jan. 2011.

[20] J. Son, A. V. Dastjerdi, R. N. Calheiros, X. Ji, Y. Yoon, and R. Buyya, "Cloudsimsdn: Modeling and simulation of software-defined cloud data centers," in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, May 2015, pp. 475–484.

[21] A. Barker, B. Varghese, J. S. Ward, and I. Sommerville, "Academic cloud computing research: Five pitfalls and five opportunities," in *Proceedings of the 6th USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 2–2.

[22] Netcraft, "Amazon web services' growth unrelenting," aug 2015, http://news.netcraft.com/archives/2013/05/20/amazon-web-services-growth-unrelenting.html.

[23] K. W. Cameron, R. Ge, and X. Feng, "Designing computational clusters for performance and power," in *Architectural Issues*, ser. Advances in Computers, M. V. Zelkowitz, Ed. Elsevier, 2007, vol. 69, pp. 89 – 153.

[24] A. Y. Zomaya and Y. C. Lee, *Energy Efficient Distributed Computing Systems*, 1st ed. Wiley-IEEE Computer Society Pr, 2012.

[25] T. Chieu, A. Mohindra, and A. Karve, "Scalability and performance of web applications in a compute cloud," in *e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on*, Oct 2011, pp. 317–323.

[26] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Gener. Comput. Syst.*, vol. 28, no. 5, pp. 755–768, May 2012.

[27] L. Ponciano and F. Brasileiro, "On the impact of energy-saving strategies in opportunistic grids," in *Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on*, Oct 2010, pp. 282–289.