Provisioning and Resource Allocation for Green Clouds

Guilherme Arthur Geronimo, Jorge Werner, Carlos Becker Westphall, Carla Merkle Westphall, Leonardo Defenti
Networks and Management Laboratory, LRG
Federal University of Santa Catarina, UFSC
Florianópolis, Brazil
E-Mail:{arthur,jorge,westphal,carla,ldefenti}@lrg.ufsc.br

***Abstract*—The aim of Green Cloud Computing is to achieve a balance between the resource consumption and quality of service. In order to achieve this objective and to maintain the flexibility of the cloud, dynamic provisioning and allocation strategies are needed to regulate the internal settings of the cloud to address oscillatory peaks of workload. In this context, we propose strategies to optimize the use of the cloud resources without decreasing the availability. This work introduces two hybrid strategies based on a distributed system management model, describes the base strategies, operation principles, tests, and presents the results. We combine existing strategies to search their benefits. To test them, we extended CloudSim to simulate the organization model upon which we were based and to implement the strategies, using this improved version to validate our solution. Achieving a consumption reduction up to 87% comparing Standard Clouds with Green Clouds, and up to 52% comparing the proposed strategy with other Green Cloud Strategy.**

***Keywords*-Green Clouds; Provisioning; Resource Allocation.**

I. INTRODUCTION

This paper proposes to improve the sustainability of Private Clouds, suggesting new strategies for provisioning and allocation for physical machines **(PMs)** and virtual machines **(VMs)**, transforming the cloud into Green Cloud and Hybrid Cloud, when needed [1]. Green Clouds crave the resource economy for the components that belongs to it. To do so, we adopt the positive characteristics of multiple existing strategies [2], developing a hybrid strategy that, in our scope, aims to address:

- A sustainable solution to mitigate peaks in environments with rapid changes and unpredictable workload.
- Optimizing the estimated data center infrastructure without compromising the availability of services, during the workload peaks.
- Improving the balance between the sustainability of infrastructure and availability of Services Layer Agreements (SLAs).

This work was based in the university data center reality, which disposes of multiple services suffering often with unexpected workload peaks, whether from attacks on servers or overuse of services in a short time.

A. Motivation

The motivation for this work can be summarized in the following points:

- **Energy saving**: Murugesan [3] says "Energy saving is just one of the motivational topics within IT environments greens.". We highlight the following points: (1) the reduction of monthly data center operating expenses (OPEX), (2) the reduction of carbon emissions into the atmosphere (depending on the country), and (3) extending the lifespan of Uninterruptible Power Supply (UPS) [4].
- **Availability of Services**: Given the recent wave of offering products, components, and elements in the form of services (*aaS), a series of pre-defined agreements between stakeholders aimed at governing the behavior of the service that will be supplied / provided is needed [5]. According to administrators in the area of information technology, the alarming factor is agreements that provide for the availability of such rates, usually 99.9% of the time or more. Thus, the question is how to provide this availability rate while consuming little power.
- **Variation Workload**: In environments with multiple services, the prediction of workload is a very complex factor. Historical data is used most to predict future needs and behaviors. However, abrupt changes are unpredictable and end up causing unavailability of the provided services. The need to find new ways to deal with these sudden changes in the workload is evident.
- **Delayed Activation**: Activation and deactivation of resources are a common technique for reduce power consumption, but the time required to complete this process is a problem that can cause some unavailability of the services provided, generating contractual fines.
- **Public Clouds**: Given the growing amount of public clouds and the development of communication methods among clouds, like Open Cloud Consortium [6], and Open Cloud Computing Interface [7], it became possible to use multiple public clouds as extensions of a single private cloud. We considered this as an alternative resource to implement new Green Cloud strategies.

 ISBN: 978-1-61208-245-5

B. Objective

Thus, we aim to propose an allocation strategy to private clouds and a provisioning strategy for Green Clouds, which suits the oscillatory workload and unexpected peaks. We will focus on finding a solution that consumes low power and generates acceptable request losses.

This paper is organized as follows: Section 2 brings the state of the art sorted by gaps found. Section 3 explains under which model the strategies were based. Section 4 presents the proposal, the idea behind the strategies, their pros and cons and where each one should be applied and not applied, tests, and the results. In Section 5, we conclude this paper and address some future works.

II. State Of The Art

About energy consumption, the paper [8] uses a Dynamic Voltage Frequency Scaling (DVFS) strategy to decrease the energy consumption in PMs used as virtualization hosts. It adapts the clock frequency of the CPUs with the real usage of the PMs. It decreases the frequency in idle nodes and increases when is needed. But the problem is that, the major energy consumption is not in the CPU, but in the other parts of the PM, so to really decrease the energy consumption you need to turn off the PMs.

The workload balance strategy for clusters in [9], tries to achieve a lower energy consumption unbalancing the cluster workload, generating idle nodes and turning off them. Extending this idea for Cloud Computing this don't work very well in cases that the cloud is fully loaded (like in Deny-of-Service attack) and the "unbalance" can not be done. This way, we saw the necessity of VMs migrations between clouds as mandatory function, to avoid this kind case.

The paper [10] tries to decrease the hosting costs in public and/or federated clouds using the costs and fines in contracts as metrics to better allocate the resources. But it limits itself in migrating VMs, between clouds, in a pool of pre-hired Clouds. This way, we foreseen that we also could considerer the resource consumption as a metric to allocate the VMs.

III. Model

The concept of combining Organization Theory and complex distributed computing environments is not new. Foster [11] already proposed the idea of virtual organizations (VOs) as a set of individuals and / or institutions defined by such sharing rules in grid computing environments. This work concludes that VOs have the potential to radically change the way we use computers to solve problems, as well as the web has changed the way of information exchange.

Following this analogy, we have a similar view: Management Systems based on the Organization Theory, providing the means to describe why / how elements of the cloud environment should behave to achieve global system objectives, which are (among others): optimum performance, reduced operating costs, appointment of dependence, service level agreements, and energy efficiency.

This organizational structures, proposed in [12], allows the network managers to understand the interaction between the Cloud elements, how their behavior is influenced in the organization, the impact of actions on macro and micro structures and vice versa, as the macro-level processes allow and restrict activities at the micro level. It aims to provide computational models to classify, predict, and understand these interactions and their influence on the environment.

Managing Cloud through the principles of the Organization Theory provides the possibility for an automatic configuration management system, since adding a new element (e.g., Virtual Machines, Physical Machines, Uninterrupted Power Supply, Air Conditioning) is just a matter of adding a new service on the Management Group.

The proposed strategies are based on a pro-active managemnt of Clouds, which is based on the distribution of responsibilities in holes, as seen in Figure 1. The responsibility of management of the cloud elements is distributed among several agents, separated in holes, and each agent controls individually, a Cloud element that suits him.

IV. Proposal

For the conscious resource provisioning of the data center, we propose a hybrid strategy that uses public cloud as an external resource used to mitigate probable Service-level Agreements (SLA) breaches due to unexpected workload peaks. In parallel, to the optimal use of local resources, we propose a strategy of dynamic reconfiguration of the VMs attributes, allocated in the data center. Given the distributed model presented in the previous section, we use the Cloud simulation tool CloudSim [14] to simulate the university data center environment. In order to simulate a distribution faithful to reality and also stressful to the infrastructure, we decided to take two distribution workloads: a real distribution, derived from monitoring requests of the institution websites, as shown in Figure 2, and another distribution derived from a mathematical oscillatory model, as shown in Figure 3. We defined these strategies with the goal of obtaining different approaches. One approach is using real environments, which would have results that intend to reflect the reality. Another approach is biased, striving for correlating the trends of the load with the results inferred.

A. Allocation

The resource allocation strategy is a proposal that introduces a composition of two other approaches: (1) the migration of VMs, which aims to focus on the processing of cloud, and (2) the Dynamic Reconfiguration of VMs, which aims to relocate dynamically the resources used by the VMs.

1) VMs Migration Strategy: This strategy aims to reduce power consumption by disabling the idle PMs of the Cloud. To induce idleness in the PMs, the VMs are migrated and

 ISBN: 978-1-61208-245-5

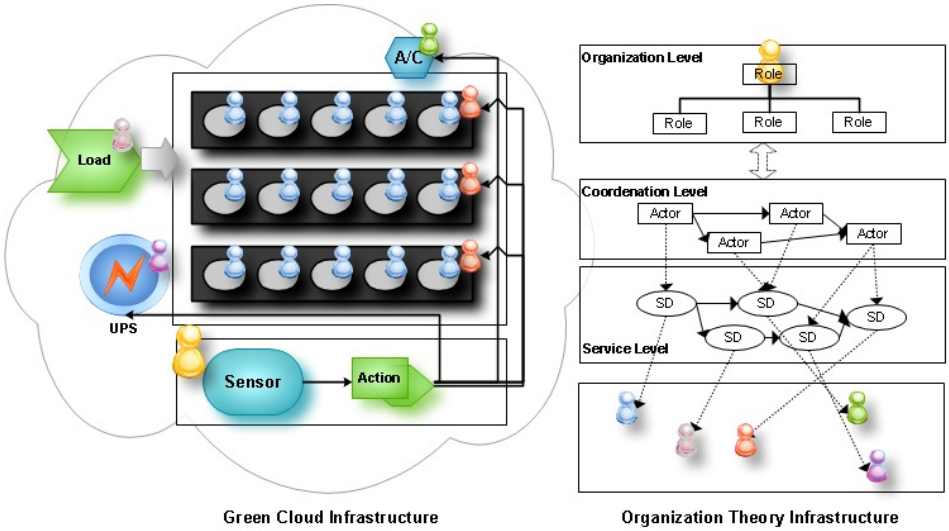

Figure 1. Model Based in Organization Theory [13]

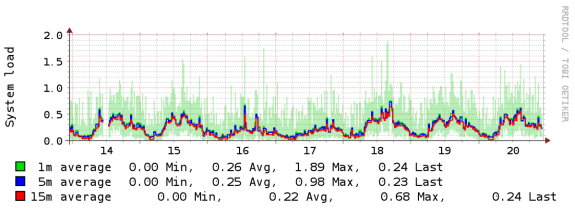

Figure 2. Real Workload Distribution

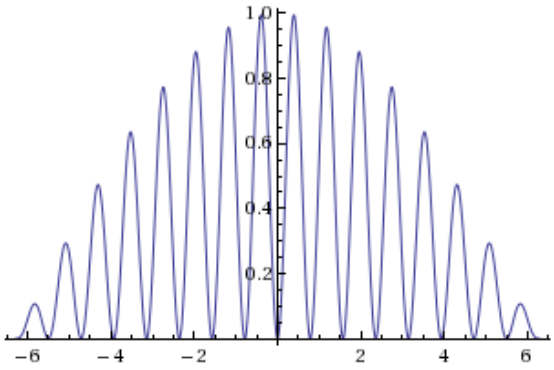

Figure 3. Oscillatory Workload Distribution

concentrated in a few PMs. This way, the cloud manager can disable the empty PMs, reducing the consumption of the data center. However, it is understood that, for the optimal results of the strategy, it must be used in conjunction with another strategy, that is a strategy that permits hosting more VMs in less PMs, generating more idle PMs.

2) VMs Dynamic Reconfiguration Strategy: Seeking the improvement of the previous strategy, this strategy is an alternative optimization to dynamically shrink the VM. It adjusts the parameters of the VM [15], without migrating it or turning it off. For example, we can increase or decrease the parameters of CPU and memory allocated. Thus, the VMs would adapt according to the demand at that moment.

3) Tests Results: To simulate the strategies we used a Cloud simulator tool developed in Melbourne, CloudSim [14]. But, in order to achieve the simulations that we need, we made some modifications in the code [13], allowing to simulate the distributions patterns and the scenario definition. Four scenarios were simulated in order to seek the comparative analysis between ordinary cloud (Scenario 1), the existing methods (Scenarios: 2 and 3), and the proposed approach (Scenario 4). Those were:

- No strategies;
- Migrating VMs Strategy;
- Reconfiguring the VMs Strategy;
- Reconfiguring and migrating VMs Strategy.

At the simulations, we gathered behavior, sustainability, and availability metrics, such as the number of idle PMs, total energy consumption, and number of SLA breaches. The graph in Figure 4 represents the energy consumption in a scenario with 100 PMs without strategies implemented. In this, the power consumption is regularly during the whole period, since all the VMs and PMs were activated during the period.

Table I shows the results of the simulations. It tells what strategies were used in each scenario and what percentage (approximate) reduction was obtained, compared to the scenario without strategies.

 ISBN: 978-1-61208-245-5

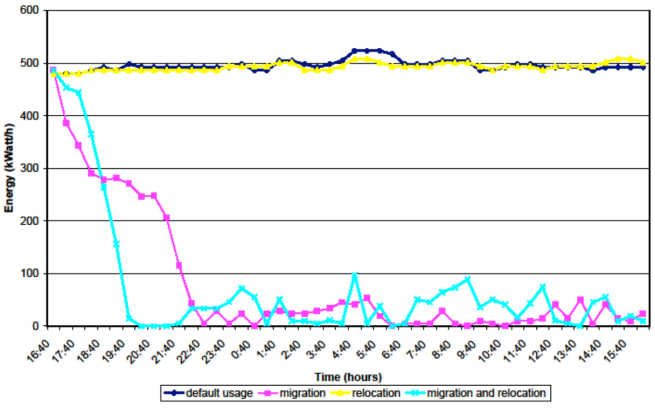

Figure 4. Energy Consumption of Scenarios 1-4 [2]

Table I
RESULTS OF ALLOCATION'S SCENARIOS

Scenario	Reconf. Strategy	Mig. Strategy	Consumption	Timeout
1	No	No	-	-
2	No	Yes	84.3%	8.0%
3	Yes	No	0.4%	-
4	Yes	Yes	87.2%	7.3%

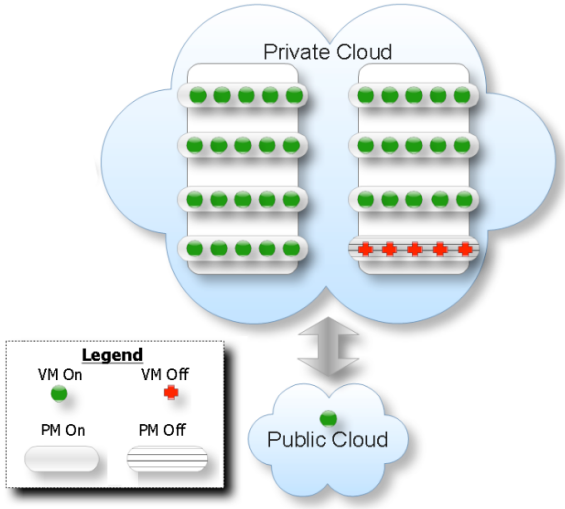

Figure 5. Hybrid Strategy [2]

B. Provisioning

The hybrid strategy is based on the merge of two other strategies, the OnDemand strategy (OD) and the Spare Resources strategy (SR). It tries to be the middle ground between the two, enjoying the strengths of both sides. It aims to present a power consumption lower than the SR strategy and a wider availability than the OD strategy.

1) On Demand Strategy: The principle of OD strategy is to activate the resources when they are needed. In our case, when a service reaches a saturation threshold, new VMs would be instantiated. When there is no more space to instantiate new VMs, new PMs would be activated to host the new VMs. The opposite also applies; when a threshold of idleness is reached, the idle VMs and PMs are disabled.

This strategy proved to be very efficient energetically, since it maintains a minimum amount of active resources. But, it has been shown ineffective in scenarios that had sudden spikes in demand, because the process to activate the resource took too much time, and the requests ended up generating losses.

2) Spare Resource Strategy: To mitigate the problem of requests timeouts, originated by a long activation time of resources, we adopt the strategy SR, whose principle is reserve idle resources ready to be used. In our case, there was always one idle VM ready to process the incoming requests and one idle PM ready to instantiate new VMs. If these resources were used, they were no longer considered idle, and new idle resources were activated. As long as the resources were no longer being used they were disabled. The strategy has been shown effective in remedying unexpected peak demands, but it showed the same behavior OD strategy in cases where demand rose very rapidly; in other words, the idle feature was not enough to process the demand. Another negative point was the energy consumption; since they always had an active and idle resource, the consumption has been greater than the OD strategy.

3) Hybrid Strategy: Seeking the merger of the strengths of the previous strategies and mitigating its shortcomings, we propose a hybrid strategy. This strategy aims to reduce the energy consumption on private cloud and reduce the breakage of SLA's service in general.

As shown in Figure 5, the cloud enables the VMs when the service in question reaches its saturation threshold, just as the OD strategy. When more PMs space is unable to allocate more VMs, it uses the public cloud to host the new VMs while the PM is passing through the activation process. This is to fulfill requests that would be lost during the activation process.

The deactivation process occurs just as the other strategies; however, it is considered that the public cloud is paid by time (usually by hour of processing); so, it disables the VM hosted in the public cloud only when (1) it's idle and (2) it is almost time to complete a full hour of hosting.

4) Tests Results: As previously mentioned, we performed some modifications to the CloudSim code, in order to enable the simulation of scenarios. Before we started the simulation, we defined some variables for the scenario, such as the saturation threshold and idleness, for example. Some of these variables are shown in Table II.

To get an overview of how each strategy would behave in different scenarios, we ran a series of tests which varied (1) the amount of requests and (2) the size of the requests.

To maintain the defined request distribution (explained in the beginning of Section 3), we used multipliers to increase

 ISBN: 978-1-61208-245-5

Table II
SIMULATION'S VARIABLES

Variable	Value
Saturation Threshold (Load 1 minute)	1.0
Idleness Threshold (Load 1 minute)	0.1
Activation VM time (seconds)	10
Activation PM time (seconds)	120
Size of Request (MI)	1000 to 2000
Number of PMs	8
Maximum number of VMs per PMs	5
SLA timeout threshold (seconds)	10

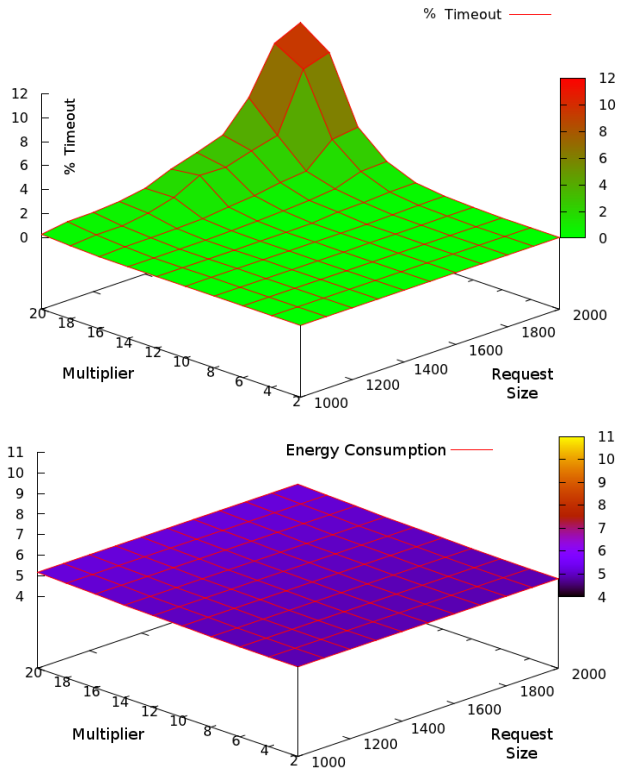

Figure 6. Number of Timeouts (top) and Energy Consumption (bottom) with Hybrid Strategy

the requests. Those multipliers started from 2 to 20 in steps of 2 (2, 4, 6, etc.).

The size of the requests ranged from 1000 to 2000 MI (Millions Instructions), in steps of 100 (1000, 1100, 1200, etc.).

This way, it performed a series of 100 simulations. This test evaluated the power consumption of the private cloud and the total number of timeouts. Figure 6 demonstrates 100 simulations in two images, the percentage of timeouts (top) and the energy consumption of the private cloud (bottom) are plotted.

Table III shows the results obtained in the "worst case scenario", by definition, with the multiplier equal to 20 and the request size equal to 2000 MI. Regarding the results in Table III, it took the Hybrid Strategy as a basis of comparison. In this case, the values listed are for hybrid strategy. For example, the hybrid strategy presented 3% less requisition timeouts than the OD strategy.

Table III
HYBRID STRATEGY COMPARED TO THE OTHER STRATEGIES

	OnDemand	Spare
Timeouts	-3 %	+15 %
Consumption	-18 %	-52 %

V. CONCLUSIONS AND FUTURE WORKS

Based on what was presented in the previous sections, and considering the objectives set at the beginning of this paper, we consider the intended goal was achieved. Two strategies for allocation and provisioning, were proposed; both aimed at optimizing the energy resource without sacrificing service availability.

The allocation strategy in private clouds, compared to a normal cloud, demonstrated a 87% reduction in energy consumption. It was observed that this strategy is not effective in scenarios where the workload is oscillating. That's because it ends up generating too much unnecessary reconfigurations and migrations. Despite this, it still shows a significant gain in energy savings when compared to a cloud without any strategy deployed.

The hybrid strategy for provisioning in green clouds, demonstrated a 52% consumption reduction over the SR strategy, and a timeout rate 3% lower than the OD strategy. Thus, we conclude that the use of this strategy is recommended in situations where the activation time of the resource is expensive for the health of SLA. We also identified that using this is not recommended when the public cloud should be used sparingly due to their course or other factors.

As future work, we aim at adding the strategy of Dynamic Reconfiguration of VMs in public clouds. This procedure was not adopted because, during the development of this work, this feature was not a market reality. We also intend to invest in new simulations of the cloud extending the variables (such as DVFS and UPS) and, if possible, explore some artificial intelligence techniques [16] such as Bayesian networks, the recalculation of beliefs. Our PCMONS (Private Cloud Monitoring System), open-source solutions for cloud monitoring and management, also will help to manage green clouds, by automating the instantiation of new resource usage [17].

We foresee, in opposition to unexpected peaks scenarios, work with cloud management based on prior knowledge of the behavior of hosted services. It is believed to be necessary to develop a description language to represent the structure and behavior of a service, enabling the exchange of information between applications for planning, provisioning, and managing the cloud.

 ISBN: 978-1-61208-245-5

REFERENCES

[1] J. Werner, G. A. Geronimo, C. B. Westphall, F. L. Koch, C. M. Westphall, R. R. Freitas, and A. Fabrin, "Aperfeiçoando a gerência de recursos para nuvens verdes," *INFONOR*, vol. 1, pp. 1–8, 2012.

[2] J. Werner, G. A. Geronimo, C. B. Westphall, F. L. Koch, R. R. Freitas, and C. M. Westphall, "Environment, services and network management for green clouds," *CLEI Electronic Journal*, vol. 15, no. 2, p. 2, 2012.

[3] S. Murugesan, "Harnessing green it: Principles and practices," *IT professional*, vol. 10, no. 1, pp. 24–33, 2008.

[4] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-Efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," in *Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010), Las Vegas, USA, July 12*, vol. 15, 2010.

[5] M. A. P. Leandro, T. J. Nascimento, D. R. dos Santos, C. M. Westphall, and C. B. Westphall, "Multi-tenancy authorization system with federated identity for cloud-based environments using shibboleth," in *ICN 2012, The Eleventh International Conference on Networks*, 2012, pp. 88–93.

[6] OpenCC, "Open cloud consortium," 2012. [Online]. Available: http://opencloudconsortium.org/

[7] OCCI, "Open cloud computing interface," 2012. [Online]. Available: http://www.occi-wg.org

[8] G. von Laszewski, L. Wang, A. Younge, and X. He, "Power-aware scheduling of virtual machines in dvfs-enabled clusters," in *Cluster Computing and Workshops, 2009. CLUSTER '09. IEEE International Conference on*, 31 2009-sept. 4 2009, pp. 1 –10.

[9] E. Pinheiro, R. Bianchini, E. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," in *Workshop on Compilers and Operating Systems for Low Power*, vol. 180. Citeseer, 2001, pp. 182–195.

[10] H. A. Franke, "Uma abordagem de acordo de nível de serviço para computação em nuvens," PPGCC/UFSC, 2010.

[11] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Environments Workshop, 2008. GCE 08*, nov. 2008, pp. 1–10.

[12] J. Werner, G. A. Geronimo, C. B. Westphall, F. L. Koch, and R. R. Freitas, "Um modelo integrado de gestão de recursos para as nuvens verdes," in *CLEI 2011*, vol. 1, 2011, pp. 1–15.

[13] Werner, J. and Geronimo, G. A. and Westphall, C. B. and Koch, F. L. and Freitas, R. R., "Simulator improvements to validate the green cloud computing approach," *LANOMS Latin American Network Operations and Management Symposium*, vol. 1, pp. 1–8, 2011.

[14] R. Buyya, "Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities," in *HPCS 2009. International Conference on*. IEEE, 2009, pp. 1–11.

[15] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Sandpiper: Black-box and gray-box resource management for virtual machines," *Comput. Netw.*, vol. 53, no. 17, pp. 2923–2938, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2009.04.014

[16] F. L. Koch and C. B. Westphall, "Decentralized network management using distributed artificial intelligence," *Journal of Network and Systems Management*, vol. 9, pp. 375–388, 2001, 10.1023/A:1012976206591. [Online]. Available: http://dx.doi.org/10.1023/A:1012976206591

[17] S. A. de Chaves, R. B. Uriarte, and C. B. Westphall, "Toward an architecture for monitoring private clouds," *Communications Magazine, IEEE*, vol. 49, no. 12, pp. 130 –137, December 2011.

 ISBN: 978-1-61208-245-5