Provisioning, Resource Allocation, and DVFS in Green Clouds

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

***Abstract*—The aim of Green Cloud Computing is to achieve a balance between 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 manage the internal settings of the Cloud, addressing oscillatory peaks of workload. In this context, we propose strategies to optimize the use of the Cloud resources while maintaining the service availability. This work introduces two hybrid strategies based on a distributed system management model; it describes the base strategies, operation principles, it validates and analyzes them, and it presents the results. In order to validate our proposed strategies, we extended CloudSim to simulate our strategies. We achieved a consumption reduction up to 87% comparing Standard Clouds with Green Clouds, up to 52% comparing the proposed strategy with other Green Cloud Strategy, and 13% less consumption using Dynamic Voltage and Frequency Scaling (DVFS) in hybrid provisioning strategy.**

***Keywords*—Green Clouds; Provisioning; Resource Allocation; DVFS.**

I. INTRODUCTION

This paper extends our previous work [1] proposing two strategies for allocation and provisioning of physical machines **(PMs)** and virtual machines **(VMs)** using DVFS as an improvement of private Clouds sustainability, transforming the Cloud into Green Cloud [2]. Green Clouds crave for efficiency of its components, so, we adopted positive characteristics of multiple existing strategies [3], developing hybrid strategies that, in our scope, aim to address:

- A sustainable solution to mitigate peaks in unpredictable workload environments with rapid changes;
- An optimization of the data center infrastructure without compromising the availability of services during the workload peaks;
- Balance between the sustainability of the infrastructure and the services availability defined on Services Level Agreements (SLAs).

This work was based on actual data collected by the university data center, that has multiple services suffering often with unexpected workload peaks, whether from attacks on servers or overuse of services in short periods of time. First, we propose an allocation model for private Clouds that aims to reduce the costs (energy and SLA fines) while improving the resource optimization. Second, we propose a provisioning model for private Clouds, turning them into Green Clouds, allowing the reduction of energy consumption and resource optimization while maintaining the Service Level Agreements (SLAs) with the integration of public Cloud resources. Third, after we validate our hybrid provisioning strategy, we have the opportunity to apply the hybrid provisioning strategy in a Cloud environment that uses Dynamic Voltage and Frequency Scaling (DVFS) in its physical machines. This way we achieve an improvement in energy consumption and resource optimization with no impacts on the Cloud SLAs.

A. Motivation

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

- **Energy saving**: Murugesan [5] says "Energy saving is just one of the motivational topics within green IT environments." 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) the extension of the lifespan of Uninterruptible Power Supply (UPS) [6].
- **Availability of Services**: Given the wave of products, components, and computing elements being delivered as services by the Cloud (*aaS), a series of pre-defined agreements or governing the behavior of the service that will be supplied / provided is needed [7]. According to Cloud Administrators, agreements that provide availability rates, usually 99.9% of the time (or more) are a concerning factor. Thus, the question is how to provide this availability rate while consuming little power.
- **Variation Workload**: In environments with multiple services, the workload prediction is complex work. Historical data is mostly used to predict future needs and behaviors. However, abrupt changes are unpredictable causing temporary unavailability of 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 reducing power

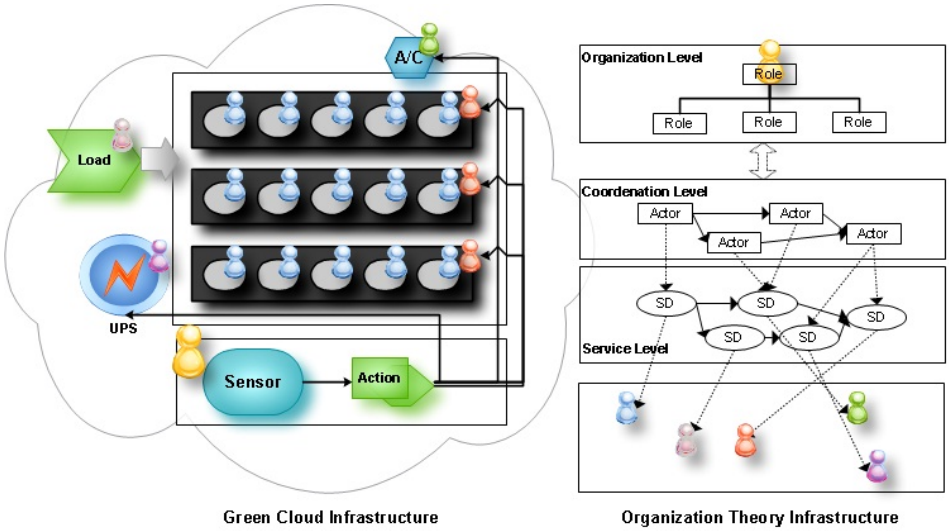

Fig. 1. Model Based in Organization Theory [4]

consumption, but the time required to complete this process can cause some unavailability of provided services, generating contractual fines.

- **Public Clouds**: Given the growing amount of public Clouds and the development of communication methods among Clouds, like Open Cloud Consortium [8], and Open Cloud Computing Interface [9], it became possible, for small or big companies, to easily use multiple public Clouds as extensions of a single private Cloud. We considered this as an alternative resource to implement new Green Cloud strategies. This is beneficial to those who need to expand their Cloud, and to the new clients of Cloud providers.

In a broad sense, this proposed model is for the Cloud provider that seeks the balance between energy saving and service providing (defined by the SLA).

B. Objective

We aim to propose an allocation strategy for 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, in comparison to other base strategies.

C. Paper Organization

This paper is organized as follows:

- Section II explains the bases of Organization Theory Model.
- Section III explains how DVFS works; this is one of the strategies used to compare our Green Cloud provisioning Hybrid Strategy.
- Section IV brings the state of the art, which includes a selection of works that were considered in our research.
- Section V presents the proposal, the idea behind each strategy, their pros and cons and where each one should be applied or not, tests, and results.
- In Section VI, we conclude this paper making some observations and analysis about the results and address some future works.

II. Model

The concept of combining Organization Theory and complex distributed computing environments is not new. Foster [10] 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 the same way as the Web has changed the way we consume and create information.

Following this analogy, we have a similar view: Management Systems based on the Organization Theory would provide means to describe why / how elements of the Cloud 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.

These organizational structures, proposed in [11], allow network managers to understand the interactions between the Cloud elements, how their behavior is influenced in the organization, the impact of actions on macro and micro structures, as the macro level processes allowing and restricting activities at the micro level. This way, it provides computational models to classify, predict, and understand the elements interactions and their influence on the whole 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 management of Clouds, which is based on the distribution of responsibilities in holes. The management responsibility of the Cloud elements is distributed among several agents; each agent controls individually a Cloud element that suits him, as seen in Fig. 1.

A. Case Study

In [3], Werner et al. proposed a model based on the Organization Theory to manage a Cloud environment using decentralized management services. They proposed agents to manage the Cloud elements, each agent managing the elements that are in its area. These agents would individually monitor and manage the elements they are responsible for, orchestrating them to fulfill the norms that are imposed to the system.

Norms are the rules or agreements used as input into the system such as SLAs, energy consumption, resource optimization, air conditioning (data center temperature), etc. They are a primitive knowledge collected from experienced administrators and are used at times when decisions need to be made. In complement to Norms, Werner et al. [3] defined believes that are empirical knowledge used to improve the decisions at management. It is the junction of the practical knowledge from the norms and empirical knowledge from historical data, derived by the system, analyzing historical data traces and correlating them with the norms that have or have not been fulfilled.

Werner et al. [3] also defined roles that the agents would assume while monitoring/managing the Cloud environments or services. The roles defined for agents that act at Cloud environment level are: VM management, server management, network management and environment management. The roles defined for agents that act at service level are: monitor element, service scheduler and service analyzer.

Based on [3], we conclude that the Organization Theory model would be applicable for managing the entities of a Cloud computing environment in a decentralized way. So far, our models apply the Organization Theory ideas as describe by Werner et al. [3], using decentralized agents to monitor and manage the Cloud entities.

III. DVFS - Dynamic Voltage and Frequency Scaling

The DVFS was presented by Magklis et al. in [12]. It provides an alternative solution to decrease power consumption by giving the possibility to the PMs to independently decrease their power dissipation, by lowering the processor clock speed and supply voltage during the idle periods of time of the processor as seen on the left side of Fig. 2.

DVFS pros:

- Adaptive Consumption: lower energy consumption by adapting the processor frequency to the workload.
- Out-of-the-box: There is no need to adapt applications or services to use it.
- Management: The user (or application) is allowed to determine when to use (or not) the solution, giving the possibility to control the CPU temperature.

DVFS cons:

- Low Performance: decreasing the CPU frequency will reduce the system performance, which is expected [14].
- Inertia of Changes: The frequency takes some time to adapt to the system's needs. So, in scenarios with high load variations, DVFS could become a problem.
- Over Changes: The rapid and constant act of 'overvolting' and 'undervolting' the processor, trying to fulfill immediately the system needs, could decrease the equipment lifetime [15].

DVFS enhancements, as seen on the right side of Fig. 2, also shows a deeper level of DVFS. The idea is to apply it at the core level, not at the processor level as a global unit. Another work is trying to decrease the gap between voltage and frequency changes. The idea is to optimize the processor and build a fast DVFS that adapts quickly to system needs, as shown in Fig. 3. Kim et al. [16] use both strategies at the same time, achieving a mark of 21 % of energy saved.

IV. State Of The Art

A. Hardware Level

According to Von Laszewski et al. [17], energy consumption is a major challenge. They use a DVFS strategy to decrease the energy consumption in PMs used as virtualization hosts. It adapts the clock frequency of the CPUs to the real usage of the PMs, decreasing the frequency in idle nodes and increasing when is needed. However, the major energy consumption is not in the CPU, but in other parts of the PM, so to really decrease the energy consumption you need to turn them off.

Gunaratne et al. [18] stated that on USA just the network interface controllers (NICs) consume hundreds of millions of US dollars in electricity every year. That amount of energy used by the NICs is growing rapidly as the default 100Mbps controllers are being replace by brand new 1Gbps controllers, which consume about 4 W more than a 100Mbps controllers. They also found out that idle and fully loaded Ethernet links consume about the same amount of power while the amount of power used by an Ethernet link is actually dependent on

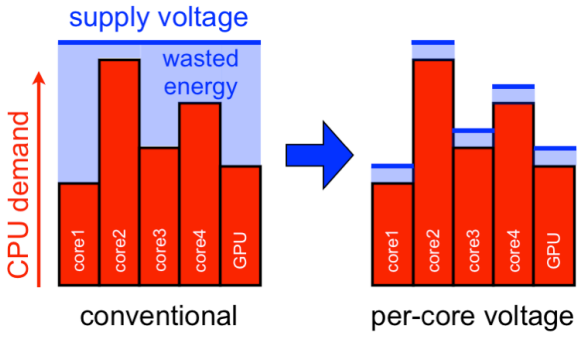

Fig. 2. DVFS - Main Idea [13]

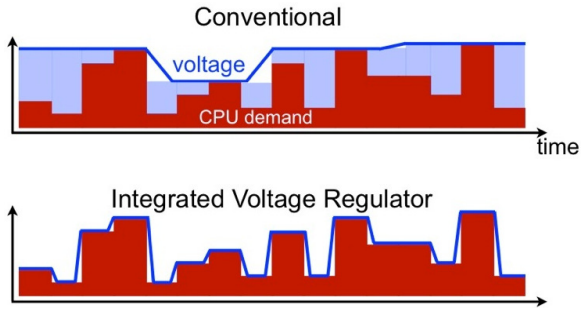

Fig. 3. Fast DVFS - Main Idea [13]

the link speed. Given the fact that measurements shown that the usage rate of Ethernets links are about 1% to 5% of the capacity, that brought attentions to the "Network Layer" as a new field to lower the energy consumption in the datacenter.

Gunaratne et al. [18] proposed a system design for adaptive link rate (ALR) [19] to be applied not just on edge links but rather on the whole network. With this approach was possible to operate Ethernet links 80% of the time on lower frequency, lowering the power consumption of the datacenter without affecting services and users. Given the fact that the network management is out of this paper scope, it was decided to not include the network infrastructure consumption in the final calculations.

B. Datacenter Level

The workload balance strategy for clusters in [20] tries to achieve a lower energy consumption unbalancing the cluster workload, generating idle nodes and turning them off. In Cloud Computing, this strategy will not work in the case of Denial-of-Service attacks. Because in that scenario all nodes will be on, and there will be none node to turn off. This way, we foresee the need for VM migration between Clouds as mandatory function, to avoid cases where the unbalance of the load cannot be done.

Urgaonkar [21] proposed an overbooking strategy to consolidate virtual machines in physical machines and this way a lower resource consumption would be achieved. But, this work did not care that much for service degradation generated by resource contention that happens when you consolidate workloads.

Anh Vu [22] proposed a model that instead of taking just historic data of workload resource consumption and behavior for resource allocation and provisioning also takes in the interference generated by applications that compete for resource. It applies a canonical correlation analysis technique to find the resources that influence the most the application behavior; this way they could consolidate workloads with less impact on provided services. Their model presents a better result than [21], but, it has high computational costs and still does not have a good performance while predicting workload needs.

Gong [23] proposed PRESS (PRedictive Elastic ReSource Scaling for Cloud systems), a lightweight model to predict workload resource needs, based mainly on historical data. It uses a Fast Fourier Transform (FFT) to spot dominant frequencies and identify workload behaviors of resource usage. When there are no dominant frequencies found it applies a Markov Chain technique to predict the workload resource need for a short period of time. It is an early work that does not have the overhead problems found in [22], but it still does not have a great performance, since there are much more variables to take in, such as background workload, VMs migration need, application design, etc.

Shen et al. [24] improved the work done in [23]. They propose a smart model for provisioning resource that aims at reducing the SLAs breaches The PRESS prediction was extended to round data values. This way, it would achieve a better result since PRESS has not been so accurate to predict workload need. In order to improve the management of resources by the predictive model PRESS, they added SLAs breaches measurement as a new variable into the prediction model. It was also proposed a predictive migration model since the migration is one of the most expensive processes when dealing with virtualized resources. It is best to start a migration before the resource contention happens, avoiding this way a long period of service degradation. They keep track of all VMs needs in a physical machine of the Cloud. This way, when the resource prediction of VMs in that physical machine uses up the amount of resources the machine has, the model triggers a migration process before the resource contention happens.

C. Cloud Level

Franke [25] tries to decrease the hosting costs in public and/or federated Clouds using costs and fines in contracts as constraints to better allocate resources. But, it limits itself in migrating VMs between Clouds, in a pool of pre-hired Clouds. This way, we foresaw that the resource consumption should also be considered as a metric to allocate the VMs.

Dawoud [26] highlights that the use of "historical resource usage traces" by themselves are not enough for a predictive model. That could lead to wrong actions at management level, especially when dealing with Web applications that usually are deployed in a multi-tier way (front end, application layer and database). So, he proposed a model to manage Web applications (in public Cloud environments) that correlates three factors, (1) historical traces of resource usage, (2) workload and (3) request types. Given the fact that the Web Applications are, in most cases, developed in a multi-tier way, the work raises the attention to the fact that each tier load does not interfere with the other tiers equally. That means, if we give more resource to a tier, we should proportionally increase the resource of every single tier of the application, since they all are tied together.

Hulkury et al. [27] proposed an integrated Green Cloud Computing Architecture that addresses the workload placement problem, determining the better place to deploy the users jobs based on their theoretical energy consumption. It requires a manager (cloud client side) to provide the jobs SLAs, job descriptions, network and server specifications, to calculate the energy consumption of the job in each cloud scenario (local, private or public Cloud). Just like [11], it touches the point of

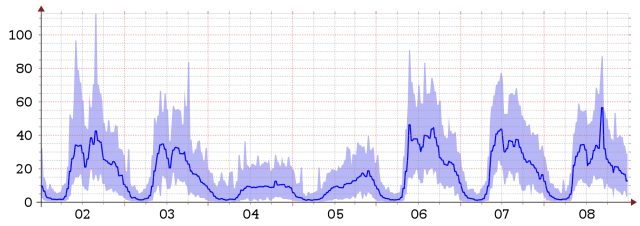

Fig. 4. Week Workload Distribution (Reqs/s)

using public clouds as an extension, and routing jobs between the clouds when it can be profitable. Sadly, it depends on some information that, in most cases, the Cloud client does not have access to, like the energy consumption of the public Cloud elements. It also mentions the idea of using XML to store SLAs and QoSs constraints in the Cloud Manager; however, it does not define any standard for that.

V. PROPOSAL

For the conscious resource provisioning in Green Cloud environments, we propose a hybrid strategy that uses public Cloud as an external resource used to mitigate SLA breaches due to unexpected workload peaks. In parallel, for 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 used the Cloud simulation tool CloudSim [28] to simulate the university data center environment and workload.

In order to simulate a distribution faithful to reality and also stressful to the infrastructure, we (1) chose as a workload pattern the distribution of requests from the university's main websites (as shown in Fig. 4), and then (2) multiplied the request load by factors between 2 and 20, in order to apply stress to the system. We defined this strategy with the goal of obtaining results that reflect the reality and, at the same time pushing the request rate, striving for correlating the workload behavior trends with the load.

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 consolidate VMs and optimize resource utilization, and (2) the Dynamic Reconfiguration of VMs, which aims to reconfigure dynamically the resources used by the VMs, increasing the consolidation factor.

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 concentrated in few PMs. This way, the Cloud manager can disable the idle PMs, reducing the consumption of the data center. However, for optimal results, this strategy must be used with a reconfiguration strategy that enables hosting more VMs in less PMs, increasing the idle PMs.

TABLE I. RESULTS OF ALLOCATION'S SCENARIOS

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

2) VMs Dynamic Reconfiguration Strategy: Seeking the improvement of the previous strategy, this strategy is an alternative optimization that dynamically shrinks the VM. It adjusts the parameters of the VM [29], 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 its configurations according to the demand.

3) Tests & Results: To simulate the strategies we used a Cloud simulator tool developed in Melbourne, CloudSim [28]. But, in order to achieve the simulations needed, we made some changes in the code [4], allowing to simulate the distributions patterns and scenario defined before. 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:

1) No strategies;
2) Migrating VMs Strategy;
3) Reconfiguring the VMs Strategy;
4) 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. Table I presents the percentage of energy consumption in each scenario with 100 PMs. Analyzing it we see that without any strategy implemented, the power consumption is stable during the whole period, since all the VMs and PMs were activated, which happens if we implement just the reconfiguration strategy by itself. The migration scenario shows a significant reduction in power consumption, since it consolidate VMs in PMs and latter disabling the idle ones. And last, the mix of migration and reconfiguration approach shows a steady noticeable reduction in power consumption since it consolidates more VMs in fewer PMs.

B. Provisioning

The Green Cloud provisioning hybrid strategy is based on two other strategies, which are the On Demand strategy (OD) and the Spare Resources strategy (SR). It tries to be the middle ground between the two, enjoying the strengths of both sides, and aiming to present power consumption lower like the OD strategy while maintaining the availability as the SR 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. And, 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. Fig. 5 shows an OD Strategy scenario, where only the needed VMs (green

circles) and PMs (white slim rectangles) are turned on. The other units (red crosses and lined rectangles) are off.

- Pros: energetically efficient since it maintains just a minimum amount of active resources.
- Cons: ineffective in scenarios that have sudden spikes in demand because the process to activate resource takes time, and some requests end up being lost.

2) Spare Resource Strategy: To mitigate the problem of requests timeouts originated by a long activation time of resources, we adopted the strategy SR, whose principle is to 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 soon as the resources were no longer being used they were disabled. Figure 6 shows a SR Strategy scenario where the Cloud keep an idle VM (golden circle) and an idle PM (vertical lined rectangle) ready to fulfill any workload.

- Pros: The strategy has been shown effective to deal with unexpected peak demands
- Cons: It showed the same behavior as OD strategy in cases where demand raised very rapidly; in other words, the idle feature was not enough to process the demand. Another negative point was the energy consumption; since it always had an idle resource, the consumption was 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 violation of SLAs.

As shown in Fig. 7, the Cloud enables the VMs when the service in question reaches its saturation threshold, just as the OD strategy. When a PM is unable to allocate more VMs, it uses the public Cloud to host the new VMs while a disabled PM is passing through the activation process. This way we fulfill requests that would be lost during the activation process.

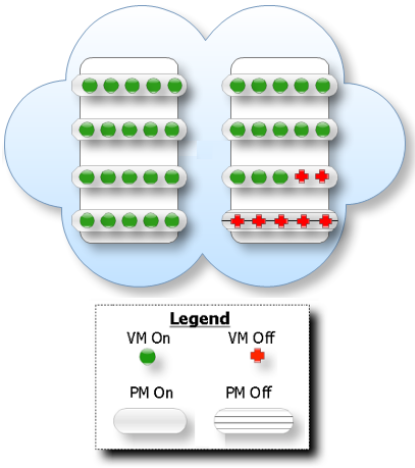

Fig. 5. On Demand Strategy

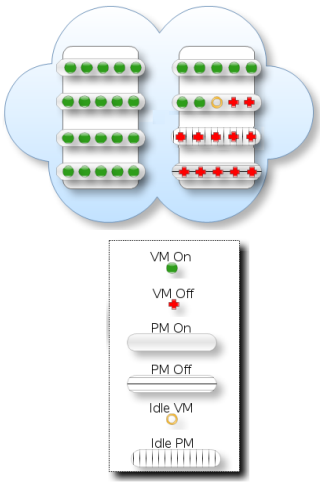

Fig. 6. Spare Resource Strategy

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:

- it is idle and;
- it is almost time to complete a full hour of hosting.

4) Tests: As previously mentioned, we performed some modifications to the CloudSim code in order to enable the simulation of scenarios using our proposed model. Before we started the simulations, we defined some variables, such as the saturation threshold and idleness. The variables considered in our experiments are shown in Table II.

The amount of requests per second was calculated based on the previously presented workload pattern (Fig. 4), using the formula $R_t * M_x$ where R_t is the number of requests per second in time *t*, and M_x is the stress multiplier of the experiment *x*.

To get an overview of how each strategy would behave in different scenarios, we ran a series of tests which varied the:

- **Amount of Requests**: To maintain the defined request

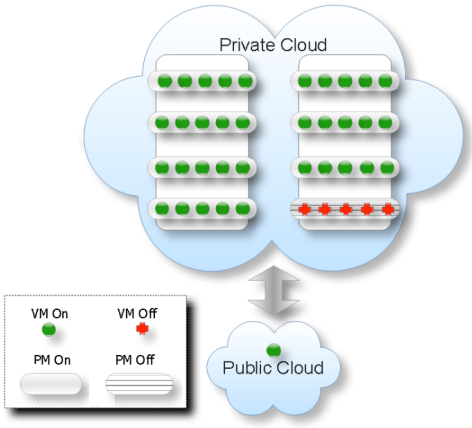

Fig. 7. Hybrid Strategy

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
DVFS	On or Off
Number of PMs	8
Maximum number of VMs per PMs	5
SLA timeout threshold (seconds)	10

distribution (explained in the beginning of Section V), we used multipliers to increase the number of requests. Those multipliers started from 2 to 20 in steps of 2 (2, 4, 6, etc.).

- **Size of Requests**: The size of requests ranged from 1000 to 2000 MI (Millions Instructions) to be executed, in steps of 100 (1000, 1100, 1200, etc.).

- **Utilization of DVFS**: Based on the previously tests, we compare the proposed hybrid strategy with and without DVFS.

This way, it was performed a total of 440 simulations being 330 simulations without DVFS and 110 with DVFS (just the hybrid strategy). These tests evaluated the power consumption of the private Cloud and the total number of timeouts (SLAs not accomplished) for the period.

5) Results: Figures 8, 9, and 10 show the results obtained while running the experiments described before. Each figure shows the timeout and the energy consumption variation of each experiment for every combination of the settings multiplier and request size variable (110 simulations each experiment).

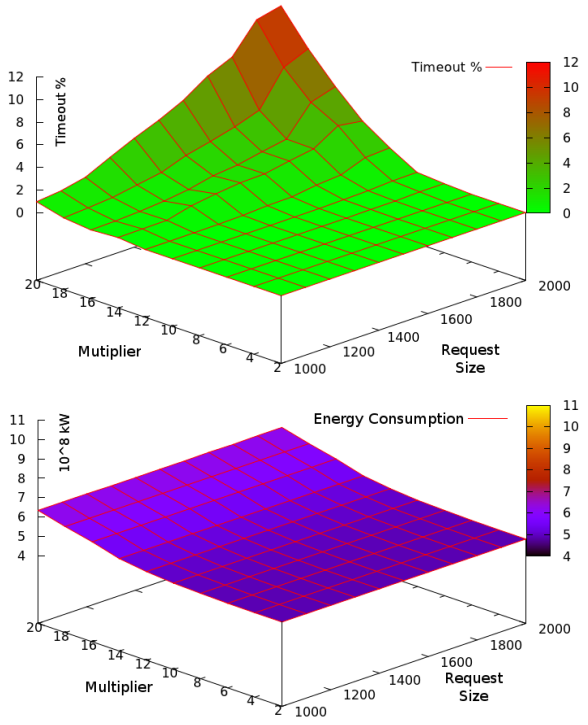

Fig. 8. Number of Timeouts (top) and Energy Consumption (bottom) using OD Strategy

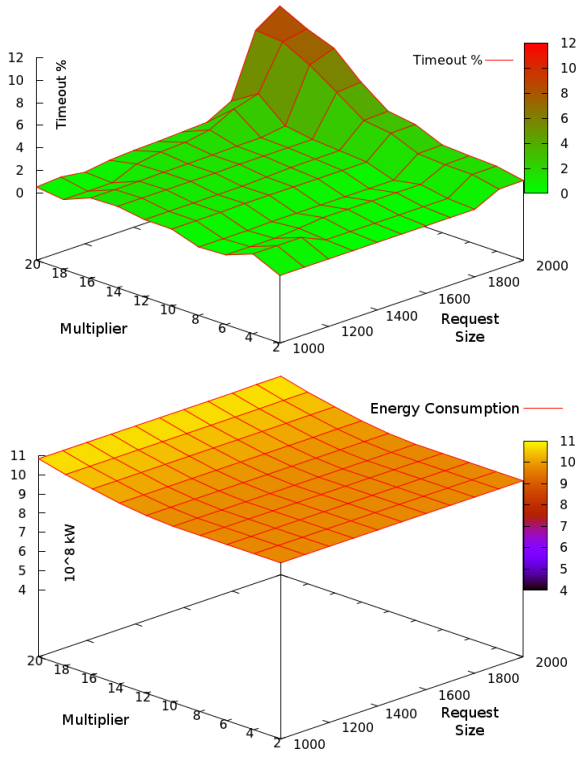

Fig. 9. Number of Timeouts (top) and Energy Consumption (bottom) using SR Strategy

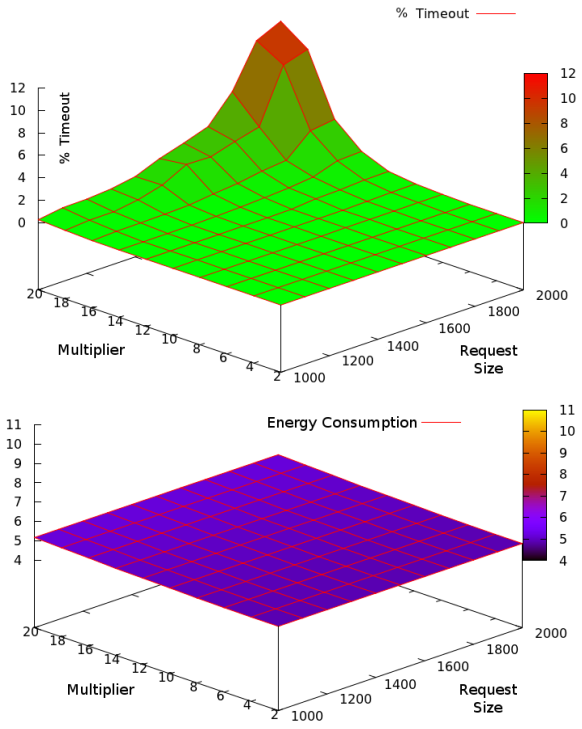

Fig. 10. Number of Timeouts (top) and Energy Consumption (bottom) using Hybrid Strategy

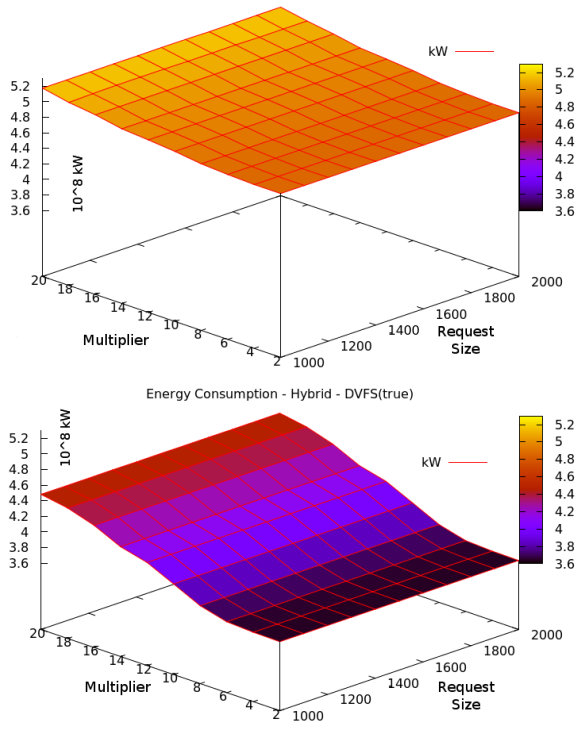

Fig. 11. Consumption with DVFS Off (top) and with DVFS On (bottom) using Hybrid Strategy

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% fewer request timeouts than the OD strategy.

TABLE III. HYBRID STRATEGY COMPARED TO THE OTHER STRATEGIES

	OnDemand	Spare	Hybrid With DVFS
Timeouts	-3 %	+15 %	-
Consumption	-18 %	-52 %	-13 %

Now, comparing the same Hybrid Strategy, with and without using DVFS, we got 13 % less energy consumption. To get a better view of the differences between the two simulations, the scale of the graph in Fig. 11 was zoomed. There were no significant difference on the timeout rate in this scenario.

VI. DISCUSSION & CONCLUSION

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

The allocation strategy in private Clouds, compared to a normal Cloud, demonstrated an 87% reduction in energy consumption. Though, it was observed that this strategy is not effective in scenarios that have huge oscillations in workload. That is because it ends up generating too much reconfigurations and migrations which have a significant computational cost. Despite this, it still shows a significative improvement in energy savings when compared to a Cloud without any resource management strategy deployed. Should be mentioned that, part of the 87% reduction rate is derived from the fact that the energy consumption from the public Cloud is not considered in the graphs. This part represents approximately 3% of the final value.

Fig. 12 shows a comparison between the green Cloud provisioning strategies. The strategies are being compared with the SR strategy which is the most expensive since it always keeps spare resources to maintain SLAs for unpredicted increases in workloads. While the OD strategy achieves up to 47.05% of energy savings when compared to SR strategy, the proposed hybrid strategy shows up to 3.13% of improvement, achieving 50.13% of energy savings with fewer timeouts than the OD strategy. The energy saving rates were even bigger when we simulated an environment where the servers deployed had the DVFS enabled. This improved the energy savings to 59.87% while maintaining the timeouts rate for extreme situations, such as when the request load was multiplied by a factor of 20 and each request size was 2000 million of instructions to be processed.

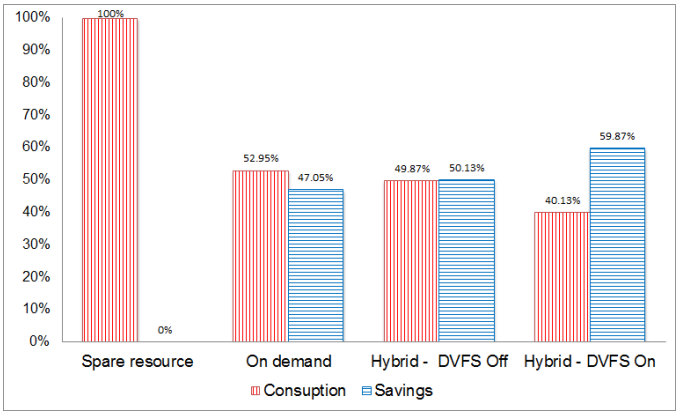

Fig. 12. Average energy consumption gain over the strategies

We should mention that we found (and fixed) some ”bugs” in CloudSim DVFS module. The simulator bases the energy consumption directly on the use of the CPU, regardless of other components in the physical machine such as GPUs, NIC, memory, HD, which leads the energy consumption to lower rates than it should be.

As can be viewed in Figures 8, 9, and 10 the strategy that achieved the lowest timeout rate was the SR strategy followed directly by the hybrid strategy with a difference lower than 3%. It was expected that the SR strategy achieved better timeout rates since it always has a spare VM and PM to supply sudden spikes in workloads, though it comes with a high cost in energy consumption and resource optimization. So, if we consider the energy savings and resource optimization generated by the hybrid strategy and compare them with the expenses of the SR strategy, the 3% extra timeouts generated by the hybrid strategy is acceptable.

Thus, we conclude that the use of the hybrid strategy is recommended in situations where the activation time of resources affects directly the SLA (in other words, generates fines). This strategy is the most balanced strategy for resource provisioning for green Cloud environments. However, this approach is not recommended when access to public Cloud resources is poor or the Cloud provider lacks in resource quality, security or other factors that can affect directly the SLAs.

A. Future Works

As future work, we aim at adding the strategy of Dynamic Reconfiguration of VMs in public Clouds. This way the public Cloud provider would be able to better manage its resource. 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 (e.g., adding UPS variable), exploring some artificial intelligence techniques [30] such as Bayesian networks, adding the recalculation of beliefs, repeating the simulation with different Cloud simulators such as GreenCloud [31], ICanCloud [32] or MDCSim [33]. This way we could compare the results and check if our proposed models show the same benefits in different simulation tools engines.

We also want to implement our proposed solutions in a real datacenter, in order to create an error factor between the results obtained with the use of simulation tools and the results of a real Cloud. This way we could measure how accurate the Cloud simulation tools are when compared with a real environment. 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 [34].

We foresee a way of working out unexpected workload peaks scenario. Prior knowledge of the behavior of hosted services could allow the management services to improve consolidation and energy consumption while maintaining the services' expected behaviors. It is believed to be necessary to develop a description language that represents the structure and behavior of a service, enabling and easing the exchange of information between application developers and Cloud provider for planning, provisioning, and managing the Cloud.

ACKNOWLEDGMENT

The authors thank anonymous reviewers for their comments to improve the paper.

REFERENCES

[1] G. Geronimo, C. Werner, J. Westphall, C. Westphall, and L. Defenti, “Provisioning and Resource Allocation for Green Clouds,” in *ICN 2013 - The Twelfth International Conference on Networks*, Jan. 2013, pp. 244–289.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] OpenCC, “Open cloud consortium,” 2012, ”[Online; Last access: 2013-01-15]”. [Online]. Available: http://opencloudconsortium.org/

[9] OCCI, “Open cloud computing interface,” 2012, ”[Online; Last access: 2013-01-15]”. [Online]. Available: http://www.occi-wg.org

[10] 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.

[11] 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.

[12] G. Magklis, G. Semeraro, D. Albonesi, S. Dropsho, S. Dwarkadas, and M. Scott, "Dynamic frequency and voltage scaling for a multiple-clock-domain microprocessor," *Micro, IEEE*, vol. 23, no. 6, pp. 62–68, 2003.

[13] W. KIM, "Fast, per-core dvfs using fully integrated voltage regulators," "[Online; Last access: 2013-01-15]". [Online]. Available: http://www.eecs.harvard.edu/ wonyoung/research.html

[14] Q. Wang, Y. Kanemasa, J. Li, C. A. Lai, M. Matsubara, and C. Pu, "Impact of dvfs on n-tier application performance," in *Conference on Timely Results in Operating Systems (TRIOS)*. ACM, 2013.

[15] M. Basoglu, M. Orshansky, and M. Erez, "Nbti-aware dvfs: A new approach to saving energy and increasing processor lifetime," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, 2010, pp. 253–258.

[16] W. Kim, M. Gupta, G.-Y. Wei, and D. Brooks, "System level analysis of fast, per-core dvfs using on-chip switching regulators," in *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*, 2008, pp. 123–134.

[17] 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.

[18] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen, "Reducing the energy consumption of ethernet with adaptive link rate (alr)," *Computers, IEEE Transactions on*, vol. 57, no. 4, pp. 448–461, 2008.

[19] C. Gunaratne, K. Christensen, and B. Nordman, "Managing energy consumption costs in desktop pcs and lan switches with proxying, split tcp connections, and scaling of link speed," *International Journal of Network Management*, vol. 15, no. 5, pp. 297–310, 2005.

[20] 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.

[21] B. Urgaonkar, P. Shenoy, and T. Roscoe, "Resource overbooking and application profiling in a shared Internet hosting platform," *ACM Transactions on Internet Technology*, vol. 9, no. 1, pp. 1–45, Feb. 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1462159.1462160

[22] A. V. Do, J. Chen, C. Wang, Y. C. Lee, A. Y. Zomaya, and B. B. Zhou, "Profiling Applications for Virtual Machine Placement in Clouds," *2011 IEEE 4th International Conference on Cloud Computing*, pp. 660–667, Jul. 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6008768

[23] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*. IEEE, 2010, pp. 9–16.

[24] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, 2011, p. 5.

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

[26] W. Dawoud, I. Takouna, and C. Meinel, "Dynamic scalability and contention prediction in public infrastructure using internet application profiling," in *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*. IEEE, 2012, pp. 208–216.

[27] M. N. Hulkury and M. R. Doomun, "Integrated green cloud computing architecture," in *Proceedings of the 2012 International Conference on Advanced Computer Science Applications and Technologies*, ser. ACSAT '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 269–274. [Online]. Available: http://dx.doi.org/10.1109/ACSAT.2012.16

[28] 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.

[29] 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

[30] 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

[31] D. Kliazovich, P. Bouvry, Y. Audzevich, and S. U. Khan, "GreenCloud: A Packet-Level Simulator of Energy-Aware Cloud Computing Data Centers," *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pp. 1–5, Dec. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5683561

[32] A. Núñez, J. L. Vázquez-Poletti, A. C. Caminero, G. G. Castañé, J. Carretero, and I. M. Llorente, "icancloud: A flexible and scalable cloud infrastructure simulator," *Journal of Grid Computing*, vol. 10, no. 1, pp. 185–209, 2012.

[33] S.-H. Lim, B. Sharma, G. Nam, E. K. Kim, and C. R. Das, "Mdcsim: A multi-tier data center simulation, platform," in *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on*. IEEE, 2009, pp. 1–9.

[34] 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.