CLEI ELECTRONIC JOURNAL, VOLUME 15, NUMBER 2, PAPER 2, AUGUST 2012

Environment, Services and Network Management for Green Clouds

Jorge Werner, Guilherme A. Geronimo, Carlos B. Westphall, Fernando L. Koch, Rafael R. Freitas, and Carla M. Westphall
Federal University of Santa Catarina, Networks and Management Laboratory,
Florianópolis, SC, Brazil, 88040-970
{jorge,arthur,westphal,koch,freitas,carla}@lrg.ufsc.br

Abstract

Green cloud computing aims at a processing infrastructure that combines flexibility, quality of services, and reduced energy utilisation. In order to achieve this objective, the management solution must regulate the internal settings to address the pressing issue of data centre over-provisioning related to the need to match the peak demand. In this context, we propose an integrated solution for environment, services and network management based on organisation model of autonomous agent components. This work introduces the system management model, analyses the system's behaviour, describes the operation principles, and presents a case study scenario and some results. We extended CloudSim to simulate the organisation model approach and implemented the migration and reallocation policies using this improved version to validate our management solution.

Keywords: Green Cloud Computing, Integrated Management, Energy Efficiency.

1 Introduction

The goal of green computing is to seamlessly integrate management of computing devices and environmental for control mechanisms to provide quality of service, robustness, and energy efficiency. The challenge in green cloud computing is to minimize resource usage and still satisfy quality of service requirements and robustness. The problem is summarised as follows. The load prediction models in traditional architectures and cloud computing environments are based on the analysis of historical data and demand increments from business models. This information makes it possible to pre-allocate resources. However, load prediction models are challenged (and frequently broken) when unexpected peaks of demand occur.

Approaches to dealing with the problems of load prediction models include the following: (i) allow for a margin of on-line resources, i.e., over-provision resources; (ii) to turn on idle resources; (iii) to temporarily use external resources on-demand (i.e., federated clouds), and others. Each of these approaches has its advantages and disadvantages. The challenge in green computing, as described by Valancius et al. [1], is to exploit the balance between these approaches in order to address the pressing issue of data centre over-provisioning related to the need to match the peak demand.

We propose a solution based on integrated environment, services and network management that promotes: (i) equitable load distribution through techniques like virtual machines; (ii) predictive resource allocation models through historical load analysis and pro-active allocation methods; (iii) aggregate energy management of network devices; (iv) integrate control over the environmental support units, which represent the larger share of energy consumption.

The objectives are the following: (i) to provide flexibility of the system configuration that allows for the easy introduction of new elements in the managed environment and the configuration processing distribution among services; (ii) to provide a level of availability that keeps to higher standard SLA (Service Level Agreement) compliance rates and which contributes to system's stability and security; (iii) to reduce cost in both capital and operational costs (CAPEX and OPEX) [2] to support the business predicates, and thus promote the acceptability of the proposed method; (iv) to provide sustainability by using methods to reduce energy utilisation and carbon emission footprints.

To achieve our objectives we propose an organisation theory model for integrated management of a green cloud computing environment. It works based on organisation models that regulate the behaviour of autonomous components (agents) that view the environmental elements, network devices (e.g. switches, cards and ports) and service providers (e.g. processing servers, load distribution services, task processors and temperature reduction

services). For example, the management system is able to turn off unused network devices and servers, turning off the environmental support units. This is reactive to characteristics of the predicted system load. The controlling elements are able to coordinate between themselves aiming at a higher-level system's objective, e.g. to keep overall energy utilisation and SLA compliance metrics.

Our research advances the state of the art as follows: (i) it introduces an organisation theory model for integrated management of the green clouds based on the concepts of organisation models, network management, and distributed computing; (ii) it analyses the network and system's behaviour and operational principles; (iii) it validates the proposal demonstrating the system's added-value in a case study scenario; (iv) it improves a simulator (the CloudSim framework) to validate the green cloud computing management approach.

This work is structured as follows. Section 2 introduces a motivating scenario. Section 3 outlines the system design. Section 4 presents a case study. Section 5 presents some results. We conclude the paper in section 6.

2 Motivation

Our research was motivated by a practical scenario at our university's data centre. In the (not so distant) past, we applied the "traditional architecture" which was composed of diverse processing clusters configured to process different services. We faced the usual issues encountered in large data centres at that time: lack of rack space, which impacted flexibility and scalability; an excessive number of (usually outdated) servers, which impacted operation costs; the need of an expensive refrigeration system; and an ineffective Uninterruptible Power Supply (UPS) system, which was problematic to scale due to the number of servers involved.

With the use of cloud computing, we managed to consolidate the number of servers using virtualisation techniques. Using this technology, we concentrated the predicted load on a few machines and kept the other servers on standby to take care of peak loads. The immediate results were very positive: reduction of rack space utilisation; lower heat emission due to the reduction in server utilisation, with consequent optimisation of the cooling infrastructure, and, a quick fix for the problematic UPS system because we had less active servers.

As part of an institutional initiative towards sustainability and eco-friendliness, our next step was to optimise energy utilisation [3] and reduce carbon emission [4]. For this, we looked at solutions from the fields of green computing and, more specifically, green cloud computing. We noticed that there was room for improvement as we consolidated resources using cloud computing. For instance, there were periods in time when the Virtual Machines (VM) were idle and the servers were underutilised. Based on the principles established by Buyya et al. [5], our goal was to promote energy-efficient management and search for methods to safely turn off unused servers using an on-demand basis. The intuitive approach was to concentrate the running applications (configured per VMs) in a few servers and recycle server capacity.

Although appealing, this approach led to a major issue: service unavailability! A quick analysis concluded that it was related to the time required to bring up the servers during unpredictable peak loads. We concluded the following: (i) the dimensioning is based on historic intra-day analysis of services demand. More specifically, it is based on the analysis of previous day's demand plus a margin of the business growth that can be estimated as the amount of resources required for one service in a period of time; (ii) however, when dealing with services with highly variable workloads, that prediction becomes complex and often immature. Moreover, external factors can lead to unexpected peaks of demand. For that, we left a safety margin of resources available (e.g. 20% extra resources on standby). Besides the excessive energy utilisation, this approach fails when the demand surpassed that threshold; (iii) as a solution, we needed to bring up turned-off resources. The lapse of time between the detection of the situation and the moment that processing resources become available caused the service unavailability.

We analysed several alternatives to overcome this issue that implements an organisation theory model for integrated management of the green clouds focusing on: (i) optimising resource allocation through predictive models; (ii) coordinating control over the multiple elements, reducing the infrastructure utilization; (iii) promoting the balance between local and remote resources; and (iv) aggregating energy management of network devices.

Our decision was in favour of these options as it addresses the core problem that is inefficient management, reliability, and cost reduction. Moreover, it contributed to improving energy efficiency and carbon emission footprints in our installations by integrating methods of environmental control management. In what follows, we detail the Green Cloud computing approach and explain how an intelligent management system fits in.

2.1 Concepts & Analysis

- Cloud computing: this structure describes the most common implementation of cloud. It is based on server virtualisation functionalities, where there is a layer that abstracts the physical resources of the servers and presents them as a set of resources to be shared by VMs. These, in turn, process the hosted services and (may) share the common resources.

- Green cloud: the green cloud is not very different from cloud computing, but it infers a concern over the structure and the social responsibility of energy consumption [6], hence aiming to ensure the infrastructure sustainability [7] without breaking contracts.
- Analysis: Table 1 relates (1) the 3 possible combinations between VMs and PMs, with (2) the average activation delay (time in seconds), and (3) the chances of the services not being processed (risk). It also presents the energy consumed (based on the work of Assuncao et al. [8]) according to each scenario.

Table 1: Relation between situations & risks & activation delay & consumption

PM State	VM State	Time	Risks	Watts	Consumption
Down	Down	30s	High	0 Ws	None
Up	Down	10s	Medium	200 Ws	Medium
Up	Up	0s	None	215 Ws	High

2.2 Related Works

Pinheiro et al. [9] have proposed a technique for managing a cluster of physical machines that minimises power consumption while maintaining the QoS level. The main technique to minimise power consumption is to adjust the load balancing system to consolidate the workload in some resources of the cluster to shut down the idle resources. This concept tries to predict the performance degradation due to throughput and workload migration based on historical trace. However, the estimated demand is static - the forecast does not consider possible fluctuation in the demand over time. At the end, besides having an economy of 20% compared to fulltime online clusters, it saves less than 6% of the whole consumption of the data centre.

Calheiros et al. [10] have developed a framework for cloud computing simulation. It has four main features: (i) it allows for modeling and instantiation of major cloud computing infrastructures, (ii) it offers a platform providing flexibility of service brokers, scheduling and allocations policies, (iii) its virtualisation engine can be customised, thus providing the capability to simulate heterogeneous clouds, and (iv) it is capable of choosing the scheduling strategies for the resources.

There is some research on cloud computing models. For example, Buyya et al. [7] suggested creating federated clouds, called Interclouds, which form a cloud computing environment to support dynamic expansion or contraction. The simulation results revealed that the availability of these federated clouds reduces the average turn-around time by more than 50%. It is shown that a significant benefit for the application's performance is obtained by using simple load migration policies.

There are some preliminary researches. For example, Buyya et al. [5] aimed to create architecture of green cloud. In the proposals some simulations are executed comparing the outcomes of proposed policies, with simulations of DVFS (Dynamic Voltage and Frequency Scaling). Their results are interesting, and they leave other possible research directions open, such as optimisation problems due to the virtual network topology, increasing response time for the migration of VMs because of the delay between servers or virtual machines when they are not located in the same data centres.

Liu et al. [6] presented the GreenCloud architecture to reduce data centre power consumption while guaranteeing the performance from user perspective. GreenCloud architecture enables comprehensive online monitoring, live virtual machine migration, and VM placement optimisation. To evaluate the efficiency and effectiveness of the proposed architecture, they used an online real-time game, Tremulous, as a VM application. Evaluation results showed that they can save up to 27% of the energy by applying GreenCloud architecture. However the adoption of a set for validation of the approach is weak. In addition managing the centralised structure is not shown.

Mahadevan et al. [11] described the challenges relating to life cycle energy management of network devices, present a sustainability analysis of these devices, and develop techniques to significantly reduce network operation power. The key insight from their network energy management experience so far is that an integrated approach which aims to minimise the total power consumed by a data centre by including network power, server power, and cooling costs as inputs to a global data centre power optimization strategy can potentially result in much greater energy savings.

2.3 Problem Scenario

To understand the problem scenario, we introduce the elements, interactions, and operation principles in green clouds. Green clouds emerged as a solution to save power by utilising server consolidation and virtualisation

technologies. Fine tuning resource utilisation can reduce power consumption, since active resources (servers, network elements, and A/C units) that are idle lead to energy waste. The target in green clouds is: how to keep resources turned off as long as possible?

The interactions and operation principles of the scenario are described below:

- There are multiple applications generating different load requirements over the day.
- A load balance system distributes the load to active servers in the processing pool.
- The resources are grouped in clusters that include servers and local environmental control units (A/C, UPS, etc.). Each server can run multiple VMs that process the requests for one specific application. Resources can be fully active (servers and VM on), partially active (servers on and VMs off), or inactive (servers and resource off). The number of servers and their status configuration is defined based on historical analysis of the load demand.
- The management system can turn on/off machines overtime, but the question is when to activate resources on-demand? In other words, taking too much delay to activate resources in response to a surge of demand (too reactive) may result in the shortage of processing power for a while. This reflects directly on the quality of service, as it could deteriorate the service availability level (even if this is a short time). On the other hand, activating more unnecessary resources causes resources to be left idle and wastes energy consumption.

3 Proposals and Solutions

Green cloud with integrated management is a structure that we see as a tendency of this area and seek like a goal. These aspects that are described below are the reference for what our model aims to fulfill. In comparison to green cloud, we infer the responsibility of consuming less energy in addition to ensuring the agreements predefined in the SLA.

- Flexibility: is state-aware of all equipment under its control, acting for when it will be necessary, not when it is needed, and plan their actions based in the information of the whole cloud. It is able to predict and execute necessary changes in hardware according to the demand of the cloud; such as slowing down an overheated CPU, turning on machines based on foreseen load coming, or triggering a remote backup in case of fire. It is able to interact automatically with public clouds [7], migrating or rising up new nodes on demand in remote clouds. It provides a better support for occasional workload peaks or DoS (Denial of Service) attacks.
- Availability: encompasses a new level by extending itself to the public clouds, allowing the creation of mirror clouds. It deals with context grouping automatically, being able to migrate these groups, or elements to public clouds.
- Cost reduction: by having an automated management based on previous experience and results, it can manage itself with minimal human intervention. It uses a 24/7 management system aiming to provide a better utilisation of the resources. It will enlarge the equipments lifetime, decrease the downtime caused by human errors and reduce the expenses by adopting smart strategies for resource utilisation. With inter-cloud communications it can adopt a minimalist configuration, ensuring local processing for most of their workload, and leaving the workload peaks to an external cloud.
- Sustainability: its structure has the ability to adopt goals for SLA, goals for energy consumption (average X kWh per day) or goals for heat emission (average Y BTU per day). The structure reacts with the environment events in order to fulfill the predefined goals. Events like UPS state down, temperature sensors accusing high degrees or fire alarms on. In parallel, adapts the environment dynamically in order to fulfill the internal goals; like decreasing the cooling system to reach consumption goals.

3.1 Organisation Theory Model

As seen in Table I, the “degree of freedom” factor of the cloud becomes narrow given the few elements (PM and VM) presented by the situations to control. We propose that breaking the centralised management service in several little management services gives us the necessary elements to increase the “degree of freedom” of the cloud, creating the possibility to achieve a balanced situation between risk and consumption.

However, with several management services in the cloud we introduce a new problem: the management of these services becomes a complex job. For this, we use the principles of organisation theory, to organise and classify such services, making them easier to control. Cloud management through the organisation theory principles gives the possibility to auto configure the management system, since the addition of a new element (such as network device, VM, PM, UPS) is just a matter of including a new service in the management group.

Hence, we propose a proactive model for cloud management based on the distribution of responsibilities for roles, shown in Figure 1. In this approach, the responsibility for managing the elements of the cloud is distributed among several agents, each one in one area. These agents will individually monitor the elements of the cloud of their responsibility. They act in an orchestrated way aiming for the fulfillment of the standards (norms).

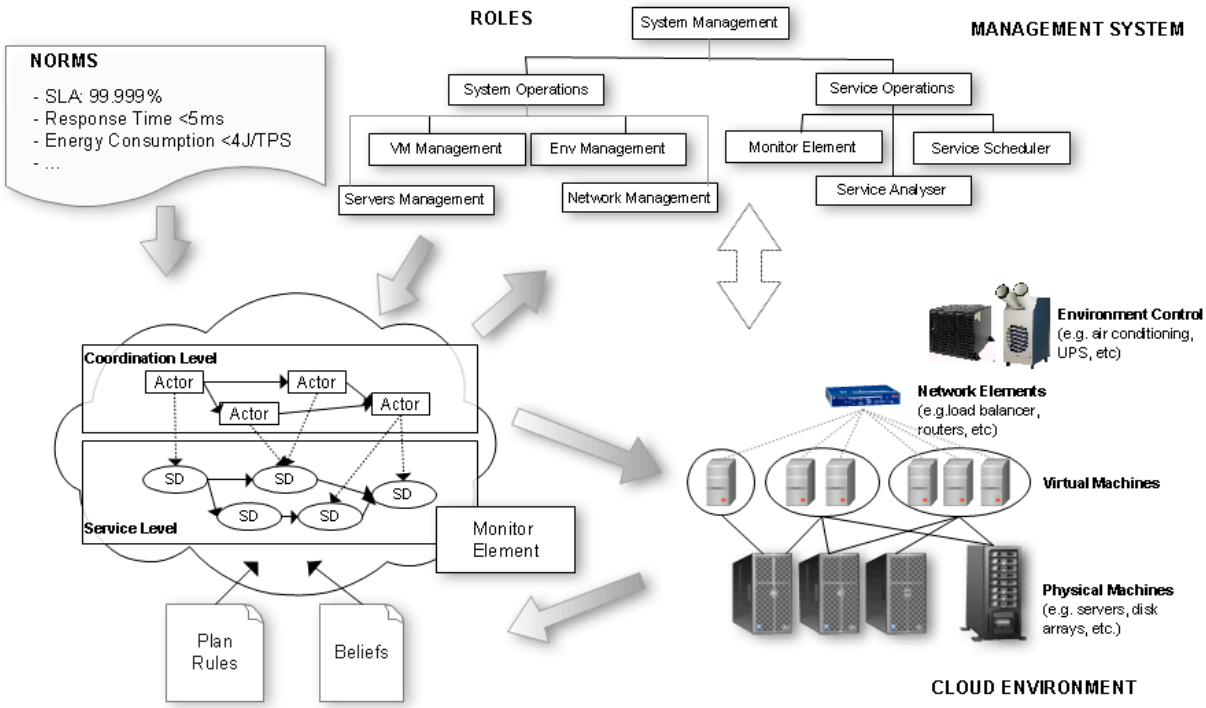

Figure 1: Integrated Management Model Proposed

Such orchestration is based on the fact that (i) the knowledge about the state of the cloud (as a whole) be shared by all agents, (ii) the existence of planning rules, to guide the actions of the agents, and (iii) the development of beliefs about the inner workflow of the cloud, that are constantly revised.

Since the data centre structure is scaled and used to provide services, this remains only a tool to provide such services. Generally, service level agreements are established in order to clarify the responsibilities of each part - client and provider. We emphasise that these agreements should be kept at their level (i.e. service), making them purely behavioural rules (e.g. delay, fault tolerance) for the service, excluding structural and physical requirements. Without the details of the environment configuration in the agreement, the cloud becomes flexible. With the independence and flexibility to change the configuration of the structure, it can become dynamic and extensible.

It can allow for covering external agreement factors still critical to the data centre infrastructure (i.e., energy consumption, hardware wear, among others), but not related to the agreement. Just as we live under the laws of physics, the cloud should also exist in well-defined laws, which we call norms. These norms express (i) the rules of the service behaviour established in the SLA and (ii) the internal interests of the cloud, which need to be considered.

For the various elements of the cloud to work efficiently, seeking the enforcement of these standards, they should be coordinated by external agents to the services they audited; managing, for example: enabling and disabling VMs; enabling and disabling PMs; configuration changes in VMs; and enabling and disabling network devices.

Since there is a wide range of elements to manage, the complexity would grow proportionally with the size of the cloud. To avoid such complexity we infer a hierarchy to the existing agents. We can make an analogy to a large company where there is a hierarchy to be followed and responsibilities being delegated. Just as in a company, there must be a system manager (the boss) that controls the entire environment. Following the hierarchy we have the

coordinators who split the operations between their teams [12] in order to facilitate the division of tasks and responsibilities among its teams.

Depending on the situation, decisions will generate system operations or service operations, or both. System operations can be divided into VM management, servers management, network management and environment management. The service operations can be divided into monitor element, service scheduler and service analyser.

The action of each role is directly reflected in the configuration of the structure as a whole. The system operations will act over the structure and environment in which the services are being processed. The services operations will act over the service layer and the environment, acquiring information from both.

3.2 Roles

The four roles that operations system may be classified as are:

- VM management: responsible for the actions implied the virtual machines. It has an interface between the model and the virtual machines. As an example, creating or destroying a VM, changing your settings and even moving it from one host to other host (either from local or remote data centre).
- Servers management: responsible for the actions implied the physical machines. It has an interface between the physical machines and the model. As an example, turning off and on a physical machine, changing the settings of the host operating system (e.g. such as BIOS - Basic Input/Output System, SMART - Self-Monitoring, Analysis, and Reporting Technology), hardware configurations (e.g. cooler and accelerometer), and backend equipment (e.g. such as storage devices, switches and site backups).
- Network management: Responsible for actions implied the network devices. It uses tools (SNMP) gathering traffic data and computing the utilization of each port on all the switches, minimising the active network components, while turning off unused switches, and disabling unused ports saving energy.
- Environment management: responsible for actions outside the structure. It has an interface between the environment and the model. As an example, temperature control of the data centre, control of power backup systems (e.g. UPS and generator), control over the accessibility of the data centre (e.g. physical security).

The three roles that service system may be classified as are:

- Monitor element: responsible for the collection of information structure in general, and your understanding. It has the responsibility to keep the model aware of the state of the cloud by monitoring the servers, VMs, network traffic and so on. It is based on specific parameters previously configured by the System Manager, such as (i) the use of a resource and its threshold notification, (ii) the availability of network links (binary data) or (iii) idleness of some element of the structure.
- Service scheduler: responsible for the cloud agenda. It has a proactive role in the model, planning the actions to be taken before the scheduled events. In an exchange of physical machines, for example, it will generate the following list of steps to be followed: test secondary UPS; enabling secondary server; and VM's migration.
- Service analyser: responsible for testing services and behavioural analysis. It has the role of auditing the service provided by the framework and understanding it. It makes sure that the service provided is in accordance with the norms to be followed, by inferring pre-established thresholds and alerting the system manager. It monitors the quality of service that is provided, and tries to relate it with the variations in the structure, finding patterns between the performance obtained and the variants elements.

3.3 Planning Rules and Beliefs

- Planning rules: the basis of theoretical knowledge, which relates contexts and objectives. They are used at times when decisions must be made, during the planning of actions. They are pieces of primitive knowledge gleaned from the experience of managers. We can take as an example of Planning Rules the following notions: (i) if a VM increases the use of page swap ping, to decrease it, we will increase memory RAM - Random Access Memory; (ii) if the physical machine presents a high load, to decrease the load, we will move the VM with more processing to another physical machine; (iii) if the data centre presents a high load, to decrease the general load, we will turn on more physical machines.

- Beliefs: empirical knowledge used to improve the decisions to be taken. In this we have the empirical understanding above the functioning of the cloud. The beliefs express the junction of practical knowledge (the premises), coming from the norms and empirical knowledge, originating from the historical data and past experiences. The beliefs must be reviewed frequently by all elements of the model, as well as the sharing of these reviews. We can take as an example of beliefs the following notions: (i) the activation of a server type X represents an increase of Y degrees in Z minutes; (ii) the activation of a VM type A increases the consumption in B kWh; (iii) the VM type A supports C requests per second.

4 Case Study

In this section we propose techniques to automatically detect the creation of data centres. We modeled the system using Norms (NM), Beliefs (BL) and Plan Rules (PR), inferring that we would need (NM) to reduce energy consumption, reduce the costs of the cloud and maintain a minimalist structure, based on a (PR) minimum of SLA violations and reduction of changes in the environment, not forgetting parameter settings (BL) of time provisioning of virtual machines.

Based on these definitions and responsibilities, the agents' sensors respond more appropriately to balance the environment. Let's consider three services (i.e. web service, backup, remote boot) running concurrently and whose charge distribution appears to be complementary. Their high peaks (i.e., variation of workload) happen at different times. Based on inferences from NM, BL and PR agents would monitor the system and determine actions dynamically.

In this proposal the agents have two solutions to the adequacy of servers and virtual machines: at a time before the peak, migrate the virtual machine to a more robust server or turn it off. Thus the system would act more dynamically and autonomically, according to the predefined requirements. Our environment is simply all the variations of workload (input), allocating and distributing services (moving/relocating) to the reduced use of resources (system output), searching environmental sustainability.

5 Results

Due to the difficulty of replicating experiments in real environments and with the goal of performing controlled and repeatable experiments, we opted to validate the proposed scenarios using simulation. For this task we used the CloudSim framework [10]; a tool developed in Java at the University of Melbourne, Australia, to simulate and model cloud-computing environments.

We extended CloudSim to simulate the organization model approach and implemented the migration and reallocation policies using this improved version (see figure 2). In this way, one can evaluate the scenario proposed in sections II, III, and IV and reuse the implemented models and controls in further simulations [13].

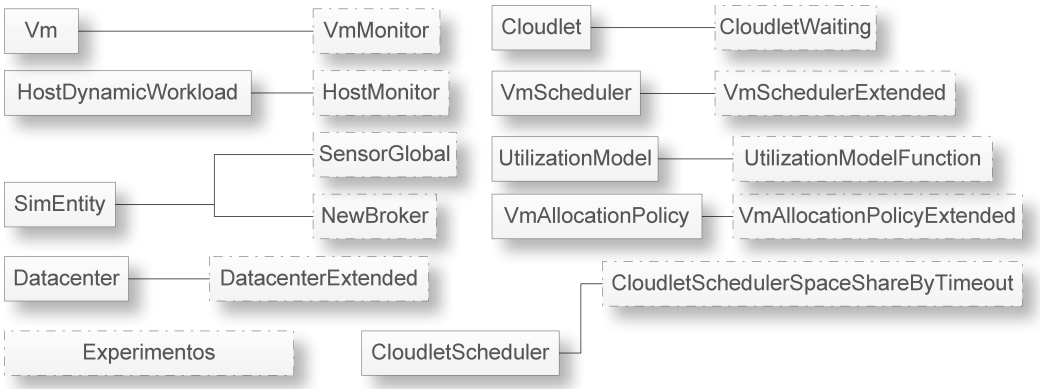

Figure 2: Classes Implemented in the CloudSim Framework [13]

We set the basic characteristics of the simulated environment, physical machines and virtual machines using data extracted from production equipments located at our university. The data was used to represent the reality of a data centre, and is based on a data centre into production at the university. It consists of different physical machines and applications that require heterogeneous virtual machine configurations. The dynamic workload was modeled according to information on peak load periods extracted from a web server. The peak load periods are random and do not present any cycle.

The main components implemented in the improved version at CloudSim are as follows:

- HostMonitor: controls the input and output of physical machines.
- VmMonitor: controls the input and output of virtual machines.
- NewBroker: controls the size of requests.
- SensorGlobal: controls the sensors.
- CloudletSchedulerSpaceShareByTimeout: controls the size and simulation time.
- VmAllocationPolicyExtended: allocation policy.
- VmSchedulerExtended: allocates the virtual machines.
- UtilizationModelFunction: checks the format of requests.
- CloudletWaiting: controls the time of the request.
- DatacentreExtended: controls the data centre.

Some experiments were simulated reaching the comparison of the usage of different VM management agents. Two kinds of agents were selected, one responsible for migrating the VM between the PM and another in charge of changing the VM configuration, like memory size or CPU frequency. Four simulations were performed, (1) one without any resource agents, (2) one applying only VM reallocating agents, (3) one applying only migrating agents and (4) one applying both agents (reallocation and migration), as presented in Figure 3.

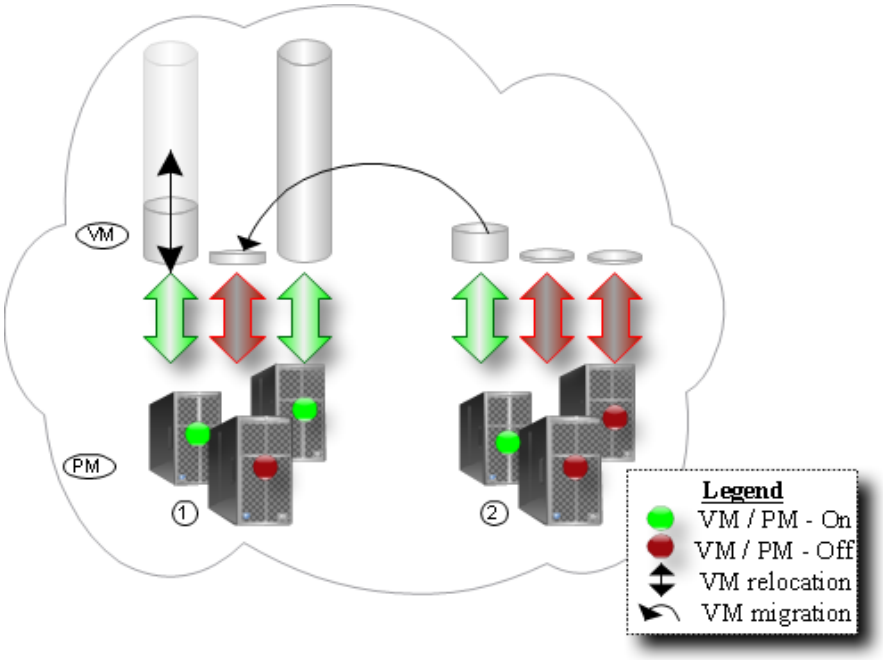

Figure 3: Simulation Using Both Policies

Figure 3 shows two clusters, each containing 100 physical machines on which virtual machines, created on demand, are allocated to applications. Some tasks are analysed, verifying the need to migrate and relocate virtual machines.

Table 2: Proposed scenario characteristics

Parameter	Value
VM - Image Size	1GB
VM – RAM	256MB
PM – Engine	Xen
PM – RAM	8GB
PM - Frequency	3.0GHz
PM – Cores	2

The simulator CloudSim was adapted to behave as the proposed model. Some parameters of the simulated scenario are presented in Table 2.

This experiment aims to verify the advantages of the strategy of using the "relocation of virtual machines" in conjunction with the strategy of "migration of virtual machines" as a resource in real time (online). It checks, analyses and addresses the changes in workload on virtual and physical machines, providing substantial resource savings in the data centre, leasing to further savings with power and air conditioning. The availability rate increases to 99.9% and the number of SLA violations decreases.

We intend to save energy by implementing policies for migrating virtual machines, allowing us to minimise the number of physical machines running. Figure 4 presents the comparison of energy consumption of the above experiments. It can be noticed a significant reduction on energy consumption of 87,18% kWatt/hour, in comparison between the experiment with ”both agents” and the experiment ”without agents” usage.

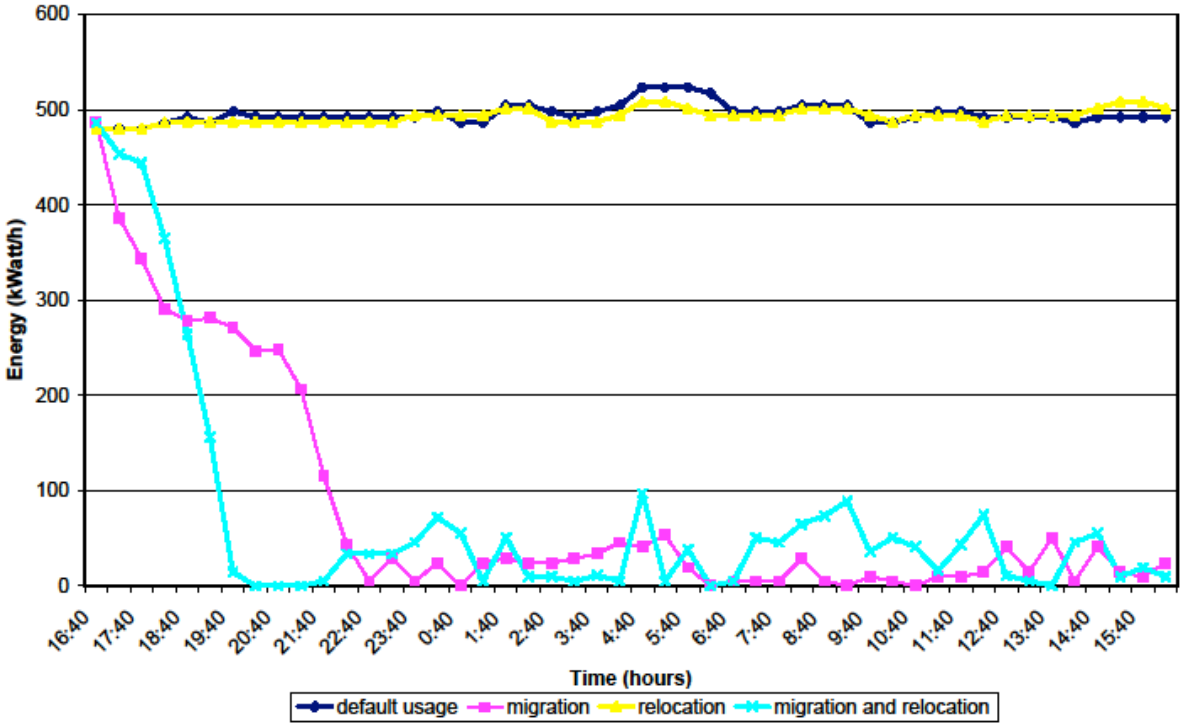

Figure 4: Intra-day Energy Consumption (kWh)

Considering the SLA without implementing policies for virtual machine migration, we have 1171 lost requests. Migrating and relocating virtual machines lead to 1077 lost requests, reducing to 8,03% the SLA violations. Figure 5 shows the SLA violations on a day.

There is a reduction in migration (45% on average over a day) an in the number of SLA violations - a result of reducing the number of lost requests. Moreover, the approach simplifies the management model, in which it is possible to manage resources (connecting / disconnecting machines) of each element and reducing energy consumption.

In a second set of experiments we simulated three allocation and activation strategies of virtual machines for green clouds. The goal was to obtain 90% of maximum workload. The three strategies are: (i) an on-demand strategy that enables physical and virtual machines when the threshold is detected; (ii) an idle resources strategy that keeps virtual and physical machines idle; and (iii) a hybrid strategy that works on demand, but in the absence of allocating physical machines to virtual machines allocates a new virtual machine in a public cloud and activates a physical machine.

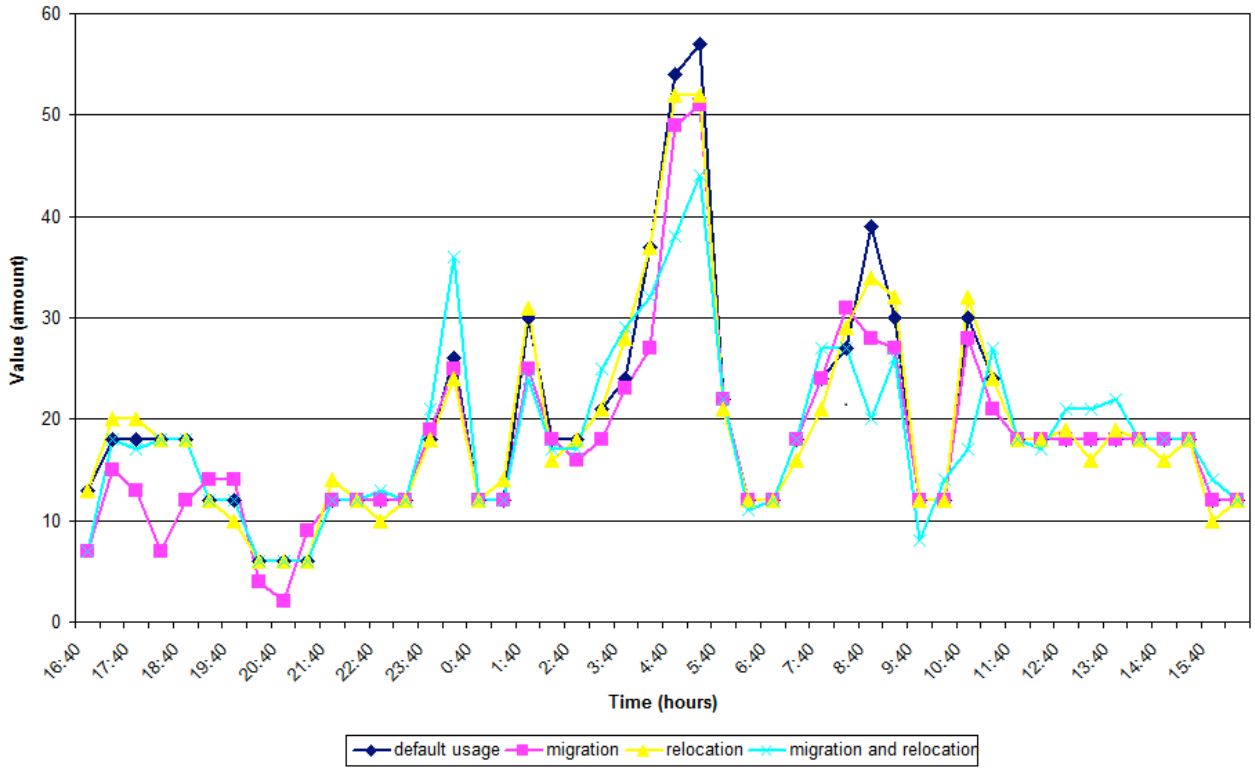

Figure 5: SLA Violations in a Day

Figure 6 shows the hybrid approach that uses a public and private cloud. The private cloud is composed of eight physical machines. The public cloud is composed of 100 physical machines. Each physical machine supports up to five virtual machines. The size of the requests was fixed at 5500 MIPS, and the maximum response time was 10 seconds.

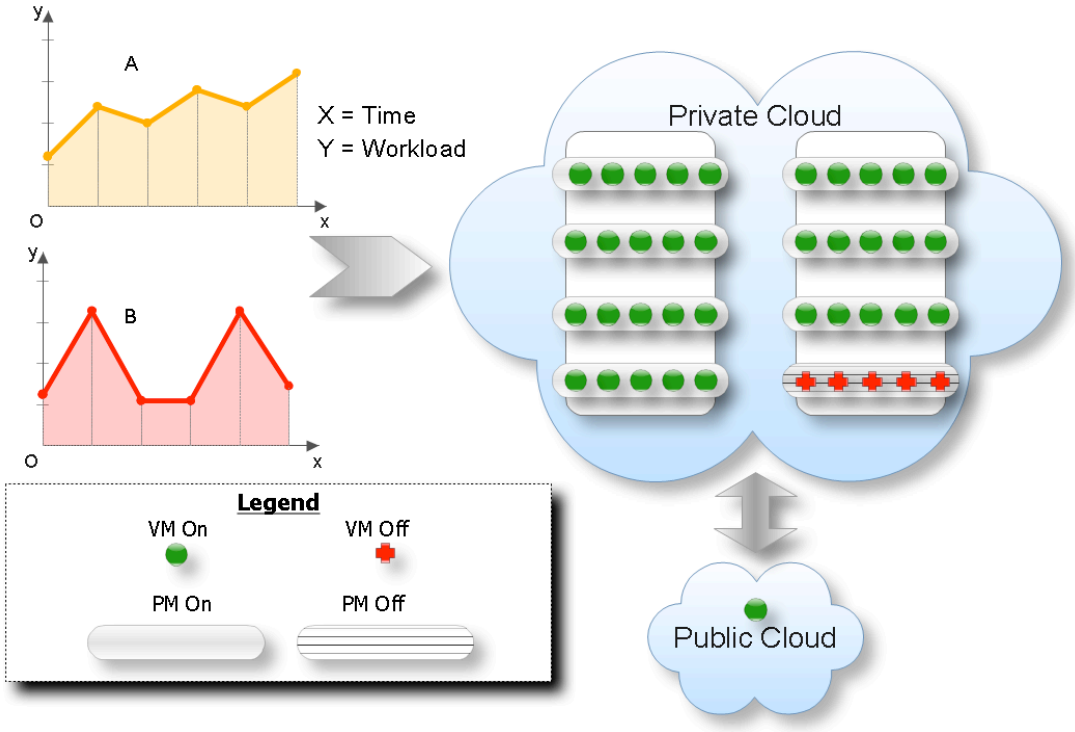

Figure 6: Hybrid Strategy for VMs Allocation

This experiment aims to verify the advantage of outsourcing the processing, using public clouds during periods of peak workload unexpected. Table 3 shows the reduction of costs and power consumption of the hybrid strategy compared to the on-demand and idle resources strategies.

Table 3: Reduction of cost and power consumption

Strategy	Cost	Consumption
On-demand	-3.2 %	-23.5%
Idle resources	-49.0%	-59.0%

Our integrated management system also has a database that stores the power constants associated with network devices (routers, switches, and line cards). Our management model predicts the power consumed by network devices during operations based on power measurements and management information, using entity Management Information Bases (MIBs) over Simple Network Management Protocol (SNMP). We have analysed the ways in which the network can be made more efficient in order to save energy, performing actions such as turning off unused switches, and disabling unused line cards and ports.

Our PCMONS (Private Cloud Monitoring System), open-source solutions for cloud monitoring and management, also helps to manage green clouds, by automating the instantiation of new or more powerful VMs, depending on the resource usage [14].

6 Conclusions and Future Works

In this paper we proposed an integrated model of environment, services and network management for green clouds based on organization model of autonomous agent components. Concepts related to cloud computing and green cloud computing were presented. We demonstrated that the proposed solution delivers both reliability and sustainability, contributing to our goal to optimize energy utilization.

Tests were realised to prove the validity of the system by utilising the CloudSim simulator from the University of Melbourne in Australia. We have implemented improvements related to service-based interaction. We implemented migration policies and relocation of virtual machines by monitoring and controlling the system. We achieved the following results in the test environment:

- Dynamic physical orchestration and service orchestration led to 87,18% energy savings, when compared to static approaches.
- Improvement in load balancing and high availability schemas provide up to 8,03% SLA error decrease.

We are building a unified power management strategy for green cloud computing, minimising the total power consumed by including network device power, server power, and cooling power.

The implementation of the experiments was successful, obtaining satisfactory results. As future work we intend to simulate other strategies to get a more accurate feedback of the model, using the simulation environment presented in [15] and testing different approaches of beliefs and plan rules. Furthermore, we would like to exploit the integration of other approaches from the field of artificial intelligence, viz. bayesian networks, advanced strategies of intention reconsideration, and improved coordination in multi-agent systems.

References

[1] V. Valancius, N. Laoutaris, L. Massoulie, C. Diot, and P. Rodriguez, "Greening the internet with nano data centers," *in CoNEXT '09: Proceedings of the 5th international conference on Emerging networking experiments and technologies*. New York, NY, USA: ACM, 2009, pp. 37–48.

[2] C. G. Gruber, "Capex and opex in aggregation and core networks," *in Optical Fiber Communication Conference. Optical Society of America*, 2009, pp. 1–3.

[3] L. Lefevre and A.-C. Orgerie, "Designing and evaluating an energy efficient cloud," *The Journal of Supercomputing*, vol. 51, pp. 352–373, 2010.

[4] C. Google, "Google data center power usage efficiency," August 2010. [Online]. Available:

http://www.google.com/corporate/datacenters/measuring.html

[5] R. Buyya, A. Beloglazov, and J. H. 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 (PDPTA2010),* Las Vegas, USA, July 12-15, 2010.

[6] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, Q. B. Wang, and Y. Chen, "Greencloud: a new architecture for green data center," *in ICAC-INDST '09: Proceedings of the 6th international conference industry session on Autonomic computing and communications industry session*. New York, NY, USA: ACM, 2009, pp. 29–38.

[7] R. Buyya, R. Ranjan, and R. Calheiros, "Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services," *in Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing*. LNCS, Springer, 2010.

[8] M. Dias de Assuncao, L. Lefevre, and A.-C. Orgerie, "Investigating the energy consumption of virtual machines," *in Energy 2010 conference: The First International Conference on Energy-Efficient Computing and Networking*, Passau, Germany.

[9] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," *in Proceedings of the Workshop on Compilers and Operating Systems for Low Power (COLP'01),* Sep 2001, pp. 182–195.

[10] 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," *Software: Practice and Experience*, vol. 41, pp. 25–50, 2011.

[11] P. Mahavadevan, S. Banerjee, P. Sharma, A. Shah, P. Ranganathan. "On Energy Efficiency for Enterprise and Data Center Networks," *in IEEE Communications Magazine*. August 2011.

[12] F. Dignum, V. Dignum, J. Padget, and J. Vazquez-Salceda, "Organizing web services to develop dynamic, flexible, distributed systems," *in iiWAS'09: Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*. New York, NY, USA: ACM, 2009, pp. 225–234.

[13] J. Werner, G. A. Geronimo, C. B. Westphall, F. L. Koch, R. R. Freitas. Simulator Improvements to Validate the Green Cloud Computing Approach," *IEEE Latin American Network Operations and Management Symposium*. 2011.

[14] S. A. Chaves, R. B. Uriarte, C. B. Westphall, "Toward an Architecture for Monitoring Private Clouds," in *IEEE Communications Magazine*, v.49, pp. 130-137, December 2011.

[15] D. Kliazovich, P. Bouvry, Y. Audzevich, and S. U. Khan, "GreenCloud: A Packet-level Simulator of Energy-aware Cloud Computing Data Centers", *in 53rd IEEE Global Communications Conference* (Globecom), 2010.