An Intelligent Scheduling Algorithm for Energy Efficiency in Cloud Environment Based on Artificial Bee Colony

Awatif Ragmani(*), Amina El Omri, Noreddine Abghour, Khalid Moussaid, Mohammed Rida
Modeling and Optimization for Mobile Services, Big data, and Cloud Computing Research Team
Faculty of Sciences of Casablanca
University Hassan II, Morocco
(*) ragmaniawatif@gmail.com

Abstract — **The development of Cloud computing applications has initiated a substantial influence on the data center expansion around the world. This expansion has led to an increase in energy consumption within the different data centers. Aware of the environmental and financial impact of the energy efficiency, several research studies have tackled the problem of the energy efficiency at the level of the hardware as well as scheduling algorithms. However, the majority of the proposed solutions such as Dynamic Voltage and Frequency Scaling (DVFS) have demonstrated a negative effect on response time where most Cloud applications required a high responsiveness. This paper suggests an energy efficiency solution for Cloud computing by applying a fast and smart algorithm. Firstly, the paper provides an evaluation of the energy efficiency within the Cloud environment through the Taguchi experience plan. Secondly, the article introduces a bee colony scheduling algorithm which aims to optimize the energy efficiency while guaranteeing an optimum response time. The validation results acquired from the GreenCloud simulator emphasize the effectiveness of the suggested scheduling analysis methodology.**

Keywords—cloud computing; scheduling; energy efficiency; performance; GreenCloud; artificial bee colony.

I. Introduction

Quite recently, the development of Cloud computing has acquired a global and crosscutting dimension. Despite some reluctance at the beginning, more and more organizations have chosen to switch a part or all of their IT services into the Cloud environment [1], [2]. This growth trend presages a radical modification in the practices and ways of understanding IT applications and hardware in the coming years. Indeed, all the orientations go into the direction of transfers of resources in memory, calculation and storage within the Cloud. This evolution would never have been possible without the technological advances in virtualization and data center administration technologies. In short, Cloud computing offers an elastic and remotely accessible environment that meets customers' needs in terms of software as well as hardware (see Fig. 1). Thus, users will only pay for consumed services instead of investing in often expensive physical equipment that will soon be obsolete, not to mention the cost of the infrastructure administration. It should be noted that the emergence of the Cloud has created new opportunities by offering the possibility of using hardware and software resources to organizations or individuals who could not afford them. In addition, data centers that host Cloud services apply drastic security standards. Though, the decision to migrate the information system of an organization remains a complex decision that must take into account several parameters including the criticality of the data. However, this growth in demand for Cloud services is causing an expansion of data centers around the world, which has a negative impact on the environment. In the last years, energy efficiency within the Cloud environment has attracted much attention from both research teams and Cloud providers. The literature on energy efficiency within data centers indicates a multitude of methodologies which could be characterized into two categories. In the first place, we identify the work that tackles the optimization of the performance of the hardware used, notably with regard to the technologies used in the servers. This first initiative brings about results in the reduction of the energy consumed, but remains insufficient due to the fact that the servers are not the only components that consume energy on the one hand, and on the other hand it is often difficult to change all servers already functional by newer servers that will consume less energy. The second approach consists in proposing load balancing and scheduling algorithms allowing the data center to run a minimum number of servers. Indeed, it has been shown that a running server consumes 70% of its maximum power. It becomes very important to implement scheduling strategies allowing virtual machine placement and migration in order to use the minimum amount of physical resources while respecting the quality of service constraints fixed in the service level agreement (SLA). This agreement sets out the commitments of suppliers and customers in terms of performance, reliability, and safety. In this paper, we study the energy efficiency in the Cloud environment based on the GreenCloud simulator and we introduce a scheduling algorithm that allows an optimal placement of virtual machines. This algorithm has been inspired by the previous research related to the artificial bee colony. The remainder of the paper includes Section 2 which summarizes the previous work that has guided this article. Section 3 presents a synthesis of the state of the art relating to the themes addressed. Section 4 introduces the problem studied as well as the results obtained during the various simulations carried out in order to

978-1-5386-1115-9/17/$31.00 ©2017 IEEE

define an algorithm for achieving energy efficiency within data centers. Section 5 concludes this study and introduces suggestions for future improvements.

II. Related Works

Previous research on energy efficiency within data centers has demonstrated the positive impact of grouping virtual machines in order to minimize the number of hosts running. This approach minimizes energy consumption but does not always guarantee compliance with service level agreement commitments. One of the first examples of energy efficiency analysis is summarized in [3] which emphasizes the importance of energy consumption in the design process of many embedded systems in real time. The authors announce the utility of variable voltage processors in decreasing the energy used by the system. Thus, they introduce a method which allows the definition of the voltage levels to be applied to a variable voltage processor. These processors use a constant priority assignment principle in order to schedule tasks. Another study [4] was interested in the evaluation of the communication network regarding the energy efficiency. The authors have introduced a planning technique, called e-STAB, which allows the organization of the traffic of applications in Cloud model. This approach is based on efficient query planning and at the same time generates a load balancing state. The research paper [5], highlights the interest of infrastructure as a service (IAAS) suppliers in optimizing energy efficiency. Thus, through their study, they present a mathematical model of the main aspects of the IAAS model as well as the definition of a CPU utilization estimator. This study underlines the fact that the simulation of use of CPU allowed an optimization of the parameters of use. Another article [6] introduces a dynamic resource provisioning approach for allocating capacity in data centers according to usage patterns followed by clients. This approach allowed the minimization of the effect of the energy optimization consumed under real-time constraints. The proposed approach is based on the concept of using user profiles when allocating resources to virtual machines. In addition, the authors of [7] highlighted the opportunities offered by the heuristic approaches in terms of energy efficiency and minimization of CO2 emissions. The authors proposed to apply heuristic algorithms to ensure the effectiveness of the virtual machine migration policies by introducing rules for self-management of the requests for researchers in order to ensure compatibility between the supply of resources and the demand within a suitable lap of time. The authors of the article [8] have announced an approach based on exploiting the history of the use of virtual machines in order to guarantee both energy efficiency and quality of service. As reported by Guzek et al. [9], the administrators of data centers have to deal with different challenges. These challenges include performance, safety, and reliability aspects. Thus, they introduce a schema for identifying resources in terms of memory, storage and networks in the Cloud environment. The authors of [10] have introduced a different scheduling approach based on the exploitation of two competing agents in a single processor. Each agent has at his disposal a list of requests to be processed by the same machine without privilege. This is reflected in the need to minimize a function that falls within the query fulfillment time of the concerned agent. The authors stated that the purpose of the proposed approach is to identify a timetable that is in line with the objectives of the two agents. Finally, the authors of the research paper [11] suggested a scheduling methodology called MinTBT-ON, of virtual machines within data centers. The proposed approach is based on the definition of a start time, an end time, a processing time and the definition of a physical machine capacity requirement. The purpose of this approach is to plan future requests according to the time ranges while respecting the limits of the physical machines. Current research on multi-objective optimization for energy efficiency [12]–[15] is focused on the completion of a satisfactory level of energy efficiency while guaranteeing optimal performance of the system. In [12], the authors have introduced a new solution to address the problem of building residential buildings with multiple and competitive objectives that include minimizing energy consumption, decreasing financial costs, and reducing environmental impacts. This work describes a multi-objective optimization model based on a harmony search algorithm (HS). The authors have defined a model that could both minimize life-cycle cost (LCC) and carbon dioxide (CO2-eq) emissions from buildings using different parameters which include design variables. Another research study [13] have examined the compatibility of multi-objective optimization solutions with the problem of optimizing energy efficiency in buildings. The article shows that there is no optimal solution for this optimization problem because of the competition of the decision criteria used. Finally, the authors summed up the strengths and weaknesses via a concrete example of the proposed solution. The research study [14] summarized existing continuous nonlinear multi-objective optimization (MOO). The article embraces the characteristics of each method by classifying these methods into three categories including non-articulate methods, a priori articulation method of preferences, and a posteriori articulation method of preferences. In addition, the authors analyzed the solutions based on genetic algorithms to identify the strengths and weaknesses of each method. In summary, the study concludes that there is no absolute superior approach, but each case has an optimal solution that depends on the type of information available in the system studied and the level of requirement. A recent paper by Tian et al. [15] suggests optimizing the plastic molding process in two steps. The first step consists in studying the parameters through simulations carried out according to the Taguchi method. This method has been applied both for the management of simulations and for the analysis of the data based on the analysis of the signal to noise ratio and the analysis of the variance ANOVA. This step made it possible to identify the most significant system parameters for the optimization of the operations of the system. The second step included the definition of a multi-objective NSGS-II algorithm.

III. Background and State of Art

A. Cloud Computing

Although opinions on the migration speed of computer systems from the traditional model to the Cloud model remain mitigated, they all agree on the advantages and opportunities of this new paradigm. Indeed, the technologies used by the Cloud are not recent but its strength comes more from its economic concept of the payment based on consumption. More and more leaders are beginning to consider switching to Cloud model. This growing interest is justified by the specificities of the Cloud computing, including elasticity that is reflected in the ability to increase and reduce the capacity of the system, automatic and instant supply of requested services, provision of application programming interfaces (APIs), and provision of billing and usage metering models based on a pay-as-you-go model [16]. One of the challenges in managing Cloud services is the inability of the provider to anticipate customers demand. In other words, a client could solicit a Cloud service for a short time while another client would need full-time resources. Thus, the design of Cloud services must take into account the different categories of users' requirements. A second critical aspect of managing Cloud platforms is scalability. In brief, Cloud applications need to support increasing demand from a given customer and enrolling new customers within the same service. In addition, enrollment in the various Cloud services on a self-service basis has the advantage of simplifying the underwriting process by avoiding a relatively long waiting period. Finally, the billing of the services used requires the application of key performance indicators (KPIs) that offer an objective view to the customer and the Cloud provider. Moreover, the Cloud must be secure and reliable to guarantee a prosperous environment for the evolution of the proposed solutions. The Cloud is based on a concept of integrated management of physical resources and IT services. The main objective of this approach is to maintain an appropriate level of service that meets performance and safety requirements. In the Cloud computing environment, the service provider uses own hosted infrastructure within its different data centers, but it can also use the Cloud infrastructure of third-party providers. One of the weaknesses of Cloud applications is its lack of customization because providers are trying to offer services that are suitable for a large class of users. Indeed, it was noted that there is a fundamental divergence between the requirements of the user and the possibilities offered by the data centers. In other words, organizations are demanding high performance, maximum security and reliability, and elasticity of near-instant services, but these organizations face relatively low budgets [1], [17]. Among the characteristics of the Cloud, the elasticity of services and self-service remain the significant points of this model. They both shorten the acquisition time of new IT resources and align the capacity of services on demand without significant investments. There are three types of Cloud service models including the infrastructure as a service (IAAS) model, the platform as a service (PAAS), and the software as a service (SAAS). These models offer services ranging from raw infrastructure to out-of-the-box applications. The second classification of the Cloud is based on its deploying model. We identify the public Cloud that groups hosted services at a third party, the private Cloud that is hosted by the organization and the hybrid Cloud that uses both previous models (see Fig.1).

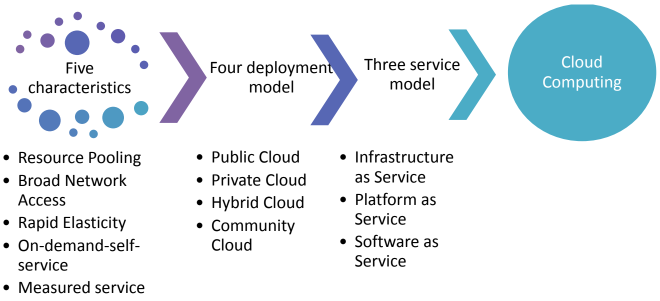

Fig. 1. The official definition of Cloud computing.

B. *Scheduling in Distributed Computing Systems*

The multiplication of networks on a large scale has been a major factor in the development of distributed computing systems. One of the fundamental aspects of distributed systems is the use of computing resources that extend over a wide geographical area. In addition, parallel processing technology has quickly gained prominence as it enables high-performance applications to be produced at lower cost. In brief, the concept of distributed processing, as well as the concept of parallel processing, remains deeply connected. Particularly, some of the distributed system approaches are applied within the framework of parallelism. Note that the distributed system is part of the class of shared memory architectures that consist of a multitude of computer nodes having different memories and clocks. In other words, the adhering nodes have individual memory and communicate via interconnection links [18]. The central role of the scheduler within a distributed architecture is the allocation of different parts of a query to multiple nodes in order to be able to handle all the queries efficiently and with better performance compared to the use of a single process. In other words, the planner aims to make the most of the physical resources by distributing workloads across multiple processors. The role of the planner within the distributed system as an administrator is organized in two steps. The first step is to cut a given program into several parallel parts. The second step is summarized in the assignment of the parts of the program to the appropriate nodes. There are two modes of scheduling. The first mode is said to be static and is characterized by the fact that the assignment of a module to a node no longer changes during the entire processing time of the module. This mode offers the advantage of minimizing the total processing time of a limited set of queries. This configuration is compatible with the use cases where the detail of the information relating to the requests to be processed is previously known. The second scheduling mode is called dynamic scheduling, which estimates a random and continuous group of incoming tasks. Unlike static scheduling, this mode does not have extensive parameters in advance. Thus, this mode is relatively more difficult to manage; however, it offers a better throughput. The implementation of an efficient scheduling policy requires taking into account a number of aspects that may influence the allocation process. These aspects include allocation parameters, static or dynamic mode selection, single or multiple query assignment, query migration policy, and load balancing within the system [18]. Indeed, load balancing plays a major role in terms of system performance. In short, load balancing is mostly done by migrating workloads from an overloaded node to a less loaded node. This technique is applied as part of scheduling. There exist several policies of load balancing and all aim at the optimization of the performance of the distributed systems. Indeed an unbalanced workload distribution will have a negative impact on performance [19]–[21]. As with scheduling policies, load balancing strategies within distributed systems consist of static solutions and dynamic solutions. In particular, dynamic load balancing embraces five steps which include estimating the load and assessing the profitability of a migration operation. The realization of an efficient load balancing passes through two phases which are the identification of the state of imbalance and the calculation of the cost. Dynamic redeployment of tasks during processing contributes to reducing the impact of load fluctuations and allows for real-time planning schedules and optimized fault tolerance of the system. Despite its many advantages, the migration mechanism remains a complex process and requires efficiently cutting the migration operation from the original environment and then transferring the work to another destination machine. This migration process must take into account the risks of communication failures by providing alternatives to avoid loss of data or performance degradation. The last aspect to take into account during the migration phase is the total cost that must be optimized in order to avoid any unfavorable impact on the system.

C. Energy Efficiency in Cloud Environment

Cloud computing is based on the use of data centers around the world. These data centers are renowned for their high energy consumption. Today the energy consumed is equal to 1.5 percent of the world's electrical power. In addition, only 15% of this energy is used for computing purposes. In short, the energy consumed within the data center is divided into the energy consumed by the computer equipment, that is to say nearly 40%, and the energy consumed by the cooling systems which is around 45% and finally the energy consumed by the energy distribution system that equals 15% of total energy. A study by the firm Gartner assessed the weight of energy consumption at 10% of the operational expenses of the data center and it envisaged that these expenses will tend to reach 50% in the years to come. This trend of growth in energy consumption has initiated several works in order to stabilize and reduce the energy consumption within the data centers. Currently, there are two concepts widely used to ensure energy saving. These concepts include dynamics and frequency cadence (DVFS) and Dynamic Power Management (DPM). The first technique relies on the dynamic adjustment of the voltage and the operating frequency according to the degree of performance required. The second technique is a control policy which focuses on accommodating the power and the system performance to its workload. In order to make the DPM technique more efficient, it should be accompanied by a scheduler that will group the queries being processed on a smaller number of servers. In addition, the scheduler must be able to evaluate the physical aspects of enabling suspended servers as well as the delays caused by this process [22]. In summary, the first energy-saving solutions include solutions for dynamically resizing servers by applying methods such as virtual machine migration, Wake-on-LAN (WoL), and dynamic frequency distribution of voltage (DVFS). These techniques nevertheless have shortcomings in terms of performance, which has motivated several researchers to propose more efficient solutions such as dynamic resource scaling (DRR), which try to minimize violations of service level agreement. This approach uses techniques that reduce the overhead caused by virtual machine allocation, server start-up, and migration of virtual machines [6].

D. GreenCloud Simulator

Regarding the complexity of distributed systems such as Cloud computing, it is often difficult to carry out detailed studies of the system parameter via a real platform. One of the solutions used by researchers is the use of simulation platforms. Particularly, the GreenCloud simulator is a proven platform. This simulator makes it possible to evaluate the energy efficiency within the Cloud environment by adapting another famous simulator which is the NS-2 platform. The GreenCloud enables the study of energy consumption at the level of the various components of data centers such as computer servers, switches, and network links. This simulator offers the possibility to program different strategies of allocation of resources and to compare their energy efficiency taking into account the network parameters. In short, the GreenCloud simulator provides the ability to perform workload distribution assessments within the data center. Particularly it allows comparing the energy efficiency of several network topologies. The GreenCloud architecture (see Fig. 2) is based on a three-level structure and can take into account the most widely used topologies such as DCell, BCube, FiConn, and DPillar. The GreenCloud architecture includes a first level that represents the root of the tree, the aggregation level that supports routing, and the level of access that groups the servers in the system. Server interconnects use 1 Gigabit (GE) Ethernet links, while the aggregation components use Gigabit connections. Thus, relying on standard racks of 48 servers, the access bandwidth is of the order of 48/20 = 2.4: 1 and of 1.5: 1 in the aggregation networks. The energy model applied by the simulator is based on the modeling of the energy consumed by the servers and the switches. At the server level, it is assumed that an inactive server uses nearly two-thirds of its maximum load while the second part of the power is consumed by the CPU according to its workload. At the switches level, the power consumption depends on the type of switch and the number of ports and cabling techniques used [22].

E. Artificial Bee Colony

In nature, several species are characterized by social behavior. We note the schools of fish, the clouds of birds, and the herds of terrestrial animals, which result from the biological necessity that drives them to live in groups. This behavior is also one of the main characteristics of social insects such ants and bees. From these principles, researchers were inspired to develop methods based on the behavior of these animals and gave birth to what is called meta-heuristics. This concept concerns all methods that model the interaction of agents that are able to self-organize. The approaches introduced by meta-heuristic allow finding out an efficient solution for the complex combinatorial problem by reiterating certain processes until finding the optimal solution. One of the most organized and rigorous insects in their work are the bees. The bees possess a great ability to communicate. The algorithm of optimization by colonies of artificial bees is a recent meta-heuristic that is inspired by the natural model of the behavior of the honey bees during their search for food. The process of searching for food by bees is based on a very efficient mechanism of displacement. This mechanism allows them to attract the attention of the bees of the colony to the food sources identified in order to collect various resources. In fact, bees use a set of wiggly dances as a mean of communication between them. These dances allow the bees to share information about the direction, distance and quantity of the nectar with its congeners [10], [23]. The collaboration and collective knowledge of the bees of the same colony are based on the exchange of information on the amount of nectar in the food source found by the different members. The bee is able, through dance or the production of chemical substances called pheromone to announce to the other bees the food source.

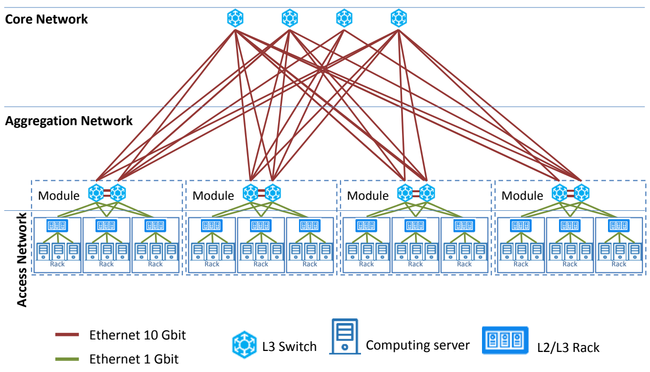

Fig. 2. The GreenCloud architecture.

The honey bee dances in circles when she found pollen at a short distance (less than 25 meters) or the bee uses a very complicated dance known as the dance in eight if the food is within 10 kilometers. The direction of food is expressed in relation to the position of the sun. The distance is expressed by the number and the speed of the turns performed by the bee on itself. In a bee colony optimization algorithm, a nectar source corresponds to a potential solution to the problem to be solved. The colony of artificial bees includes three types of bees which are the workers, the spectators, and the scouts. Firstly, the workers exploit the source of food found. They rely on their memories and try to make changes to their current position to find out a better source of food. Secondly, the spectators await the return of the workers to the dance field to observe their dances and gather information about the sources of nectar they found. Lastly, the scout bees exploit the research space by launching a random search for a new food source [24]. The initial solution population consists of an N_{FS} number of food sources randomly generated in the search space. After initialization, the solution population is subjected to repeated cycles. These cycles represent research processes performed by active, inactive scavengers and scouts. Assuming that the ith food source of the population is represented by $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$ where m is the dimension of the problem. The active beekeepers search in the vicinity of the previous source x_i for new sources v_i having more nectar, and then they calculate their fitness. In order to produce a new food source from the older one, the following expression is used:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{1}$$

$Where\ j = 1, \dots, m \quad and \quad i = 1, \dots, N_{FS}$

Despite the fact that k is determined randomly, it must be different from i. ϕ_{ij} is a random number belonging to the interval [-1, 1], it controls the production of a food source in the vicinity of x_{ij}. After the discovery of each new food source v_{ij} a gourmet selection mechanism is adopted, that is to say, that this source is evaluated by the artificial bees, its performance is compared with that of x_{ij}. If the nectar of this source is equal to or better than that of the preceding source, the latter is replaced by the new one. Otherwise, the older one is retained. For a minimization problem, fitness is calculated according to the following formula:

$$fit_i(\vec{x_i}) = \begin{cases} \dfrac{1}{1 + f_i(\vec{x_i})} & if\ f_i(\vec{x_i}) \geq 0 \\ 1 + abs\left(f_i(\vec{x_i})\right) & if\ f_i(\vec{x_i}) < 0 \end{cases} \tag{2}$$

Where $f_i(\vec{x_i})$ is the function of $\vec{x_i}$

At this stage, the inactive foragers and scouts that are expected to wait within the hive. At the end of the investigation process, the active beekeepers share the information on the nectar of the food sources and their locations with the other bees via the wiggly dance. The latter evaluate this information from all the active foragers and select the food sources as a function of the probability value P_i associated with this source and calculated by the following formula:

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \tag{3}$$

Where fit_i is the fitness of solution i, which is proportional to the quantity of the nectar of the food source of position i. The source of food whose nectar is abandoned by the bees, the scouts replace it with a new source. If during a predetermined cycle number called limit a position cannot be improved, then this source of food is assumed to be abandoned.

IV. Problem Analysis and Proposed Solution

A. Problem Statement

Given the weight of energy costs recorded by data centers compared to operating expenses combined with the environmental impact that any increase in energy consumption could have, Cloud providers tend to focus on green computing. Indeed, all Cloud providers and research community are doing their utmost to optimize the energy efficiency of Cloud infrastructures. Among these research areas, we highlight the development of a new virtual machine allocation algorithm that minimizes cost functions defined on the basis of strategic objectives such as optimal response time and very low energy consumption. However, most of the algorithms studied produce results that could be improved. Indeed, in a Cloud environment that receives n requests and which has m hosts that meet the criteria for capabilities, there are n^m way to place the queries, which makes any exhaustive search almost impossible. In addition, optimal solutions must take into account several objectives including, short response time, low cost, and very low energy consumption. Thus, approximate optimization via heuristic algorithms seems to be an efficient solution to resource allocation problems in the Cloud environment [8]. The purpose of this paper is to define an effective strategy for allocating virtual machines within the Cloud environment while respecting both the response time and energy consumption constraints. In other words, the proposed solution tends to guarantee an optimal quality of service and an economic energy system. The scheduling algorithm must take into account several aspects including the identification of the appropriate hosts by minimizing the number of hosts used while avoiding overloading certain nodes to prevent any deterioration in the system performance. In order to optimize the definition of KPIs and the identification of the factors that highly impact energy consumption, it was essential to carry out a deep evaluation of the system using a methodology defined in a previous article [25]. This method enabled us to evaluate several KPIs via the GreenCloud simulation platform. In summary, the Cloud service provider must meet both user expectations and optimize the use of resources within data centers. Achieving this balance remains fundamental in the case of real-time Cloud applications.

Particularly, the analysis of the energy consumption must be done by type of components and load of use. Based on the GreenCloud architecture, three groups of switches and a group of servers are identified. Each of those groups has an indicator for calculating the energy consumed by said group. Especially, it has been shown that an inactive server consumes nearly two-thirds of its maximum load [26]. In addition, the power of a server can vary over time. Thus, we retain as a definition of the energy consumed by the servers the following formula:

$$E_{Total} = \int_{t_1}^{t_2} k \times P_{max} + (1-k) \times P_{max} \times u(t)dt \qquad (4)$$

Where P_{max} the maximum power consumed when the server is fully used and k is equal to 70% and $u(t)$ is the workload of a server at time t. Finally, one of the indicators proposed by the American organization Green Grid for the evaluation of the energy efficiency within the data center is the power usage effectiveness (PUE). This indicator is calculated using the formula (5). Thus, a PUE ratio of 1 means that all the energy consumed by the data center is used by the IT equipment. In general, this ratio is close to 1.5 [27].

$$PUE = \frac{Total\ data\ center\ energy\ consumption}{Total\ IT\ Equipment\ energy\ consumption} \qquad (5)$$

B. Performance Analysis Methodology

The first stage of our study concerns the analysis of the factors that influence the energy consumption in Cloud environment. In this paragraph, we refer to our earlier work [25]. This method relies on the Taguchi concept to model a complex system by studying input factors and KPIs outputs. In our case study, 16 experiments have been achieved which represent 16 different system configurations. The choice of simulation scenarios has been achieved by applying the Taguchi concept. The genesis of our approach is to translate a complex system into a simpler system by evaluating only the inputs that symbolize the most influential factors and outputs that explain the key performance indicators (see Fig.3). Each input factor can take the values set by level (see Table I-II). These values may be quantitative or qualitative. The concept of Taguchi allows us to minimize the number of experiments to be carried out while maintaining the same quality of results and interpretations [28]. The Taguchi method is based on the calculation of signal-to-noise ratio (SNR) in order to rank the factors according to their effect. The optimization of the KPIs is obtained by minimizing the function (6).

$$\mathrm{SNR_i} = -10\log\left(\sum_{u=1}^{\mathrm{N_i}} \frac{\mathrm{y_u^2}}{\mathrm{N_i}}\right) \qquad (6)$$

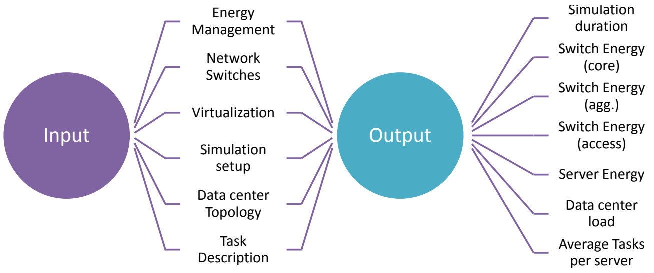

Fig. 3. Inputs and outputs of GreenCloud model.

Where i: experiment number; u: trial number; Ni: Number of trials for experiment and yu: performance representative measurements per trial.

C. Simulation and Results

During a previous paper [29], we have introduced a three-tiered architecture including different algorithms related to scheduling and load balancing in order to guarantee a high level of performance in the Cloud environment. This architecture includes several algorithms such ant colony optimization algorithms and MapReduce. Through the present paper, we focus on improving the previous solution by introducing a scheduling algorithm for energy efficiency. The purposes of the present algorithm are to ensure the energy efficiency and highest quality of service. In order to optimize the definition of the algorithm, we evaluate firstly the internal process of the Cloud environment. Regarding the complexity of carrying out repetitive and structured experiments within a real Cloud environment, we use the GreenCloud simulation platform. This simulator shares the same three-tier architecture as our proposed architecture (see Fig. 4). The simulations carried out via the GreenCloud platform allowed us to highlight the results presented in Table III. Thus, we were able to appreciate the different values of the predefined KPIs which include switch energy (core) and server Energy. The influential factors include task size, task memory, and task storage. As illustrated in Table III, except the scenario 4, the energy consumed by the servers represents the major part of the total energy consumed by a data center. In addition, scenario 9 is the configuration that consumed the most energy. In conclusion, it is obvious that any optimization of the use of the servers within the data center will have a consequent impact on the energy consumed. The regression equation of server energy is illustrated in the formula (7). This formula has oriented the definition of the fitness equation used in the proposed bee colony algorithm. According to the Tables IV-V and Fig.5, the indicator that has the greatest impact on energy consumed by switches is the core switches factor, followed by access to host while the factor that has the most influence on the energy consumed by servers' indicator is access to a host.

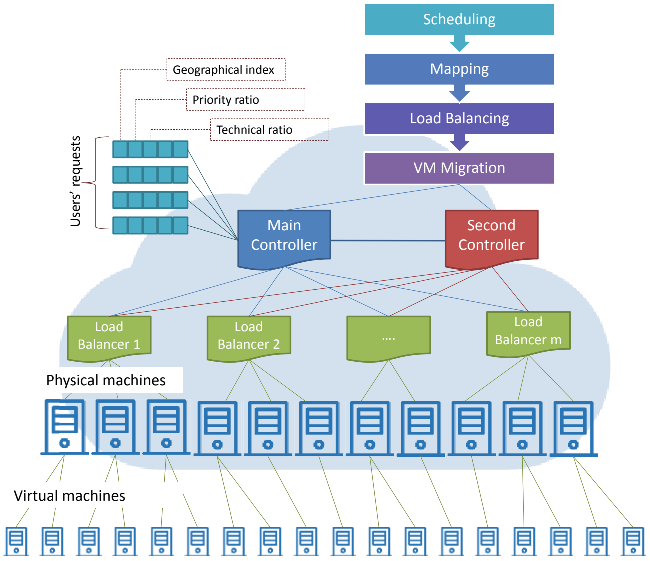

Fig. 4. The proposed Cloud computing architecture.

$$Server\ Energy = -1226.15 + 266.22\,A + 148.11\,B + 75.95C + 0.95D + 3.04E - 33.02\,F - 2.84\,G + 0.91\,H + 15.91\,I - 25.36\,J - 31.15\,K + 25.36\,L \qquad (7)$$

TABLE I. L16 TAGUCHI EXPERIENCE PLAN

Scenarios	A	B	C	D	E	F	G	H	I	J	K	L
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2
4	1	1	1	2	2	2	2	2	2	2	2	1
5	1	2	2	1	1	2	2	1	1	2	2	1
6	1	2	2	1	1	2	2	2	2	1	1	2
7	1	2	2	2	2	2	1	1	1	2	2	2
8	1	2	2	2	2	2	1	2	2	1	1	1
9	2	1	2	1	2	2	2	1	2	1	2	1
10	2	1	2	1	2	2	2	2	1	2	1	2
11	2	1	2	2	1	2	1	1	2	1	2	2
12	2	1	2	2	1	2	1	2	1	2	1	1
13	2	2	1	1	2	2	1	1	2	2	1	1
14	2	2	1	1	2	2	1	2	1	1	2	2
15	2	2	1	2	1	2	2	1	2	2	1	2
16	2	2	1	2	1	2	2	2	1	1	2	1

TABLE II. PARAMETERS VALUES PER FACTORS LEVEL

Level	Core switches	Access switches per pod	Servers per rack	Core to aggregation	Aggregation to access	Access to host
	A	B	C	D	E	F
1	5	5	12	10	1	1
2	15	10	28	100	10	10
Level	Task size (MIPS)	Task memory	Task storage	Task description size (Bytes)	Task execution deadline (s)	Task output (Bytes)
	G	H	I	J	K	L
1	300000	1000000	300000	8500	2	200000
2	600000	3000000	500000	10000	5	350000

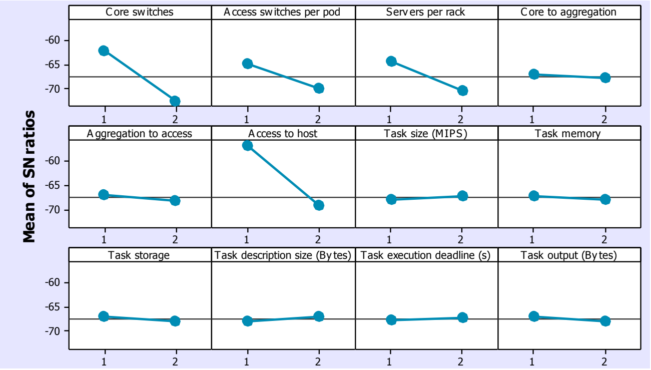

Fig. 5. Signal-to-noise ratio for Server Energy indicator.

The best energy consumption of servers could be achieved by applying the combination A1B1C1D1E1F1G2H1I1J2K2L1 which corresponds to the highest point for each factor.

TABLE III. RESULTS OBTAINED PER SCENARIO

Scenario	Simulation duration	Switch Energy (core) W*h	Switch Energy (agg.) W*h	Switch Energy (access) W*h	Server Energy W*h	Data center Load	Average Task/per server
1	62.5	263.9	527.8	65.7	664.7	26.9	223.7
2	65.5	276.6	553.1	68.9	689.6	25.6	223.7
3	62.5	263.9	527.8	65.7	664.6	26.9	111.8
4	65.5	276.6	5 531.0	68.9	689.5	25.6	111.8
5	65.5	276.6	553.1	137.8	1 381.9	25.9	113.0
6	62.5	263.9	527.8	137.8	3 119.0	27.6	115.0
7	65.5	276.6	553.1	129.5	2 911.1	26.3	229.7
8	62.5	263.9	527.8	137.3	3 119.3	27.6	229.9
9	65.5	977.0	1 954.1	215.8	4 853.6	26.3	115.1
10	62.5	932.3	1 864.6	206.0	4 679.6	27.7	115.1
11	62.5	932.3	1 864.6	206.0	4 679.6	27.7	230.2
12	62.5	932.3	1 864.6	394.4	4 005.2	27.5	228.6
13	62.5	932.3	1 864.6	394.4	4 005.2	27.5	228.6
14	65.5	977.0	1 954.1	413.4	4 154.9	26.2	228.6
15	62.5	932.3	1 864.6	394.4	4 004.9	27.5	114.3
16	65.5	977.0	1 954.1	413.0	4 154.6	26.2	114.3

TABLE IV. FACTORS RANKS FOR SWITCH ENERGY (CORE)

Factor	A	B	C	D	E	F
Rank	1	9.5	9.5	6.5	6.5	2
Factor	G	H	I	J	K	L
Rank	6.5	11.5	11.5	4	3	6.5

TABLE V. FACTORS RANKS FOR SERVER ENERGY

Factor	A	B	C	D	E	F
Rank	2	4	3	11	7	1
Factor	G	H	I	J	K	L
Rank	10	9	5	6	12	8

D. Proposed Algorithm

Based on the observations and result of the simulation stage, we introduce the scheduling algorithm I by applying the artificial bees' colony concept. Due to its remarkable capacities of convergence, the proposed algorithm focuses on identifying an optimal host for a given virtual machine in a very short time which could achieve a suitable energy efficiency level. After different configurations, the best results have been obtained by applying an acceleration coefficient equal to 0.7 (see Fig.6).

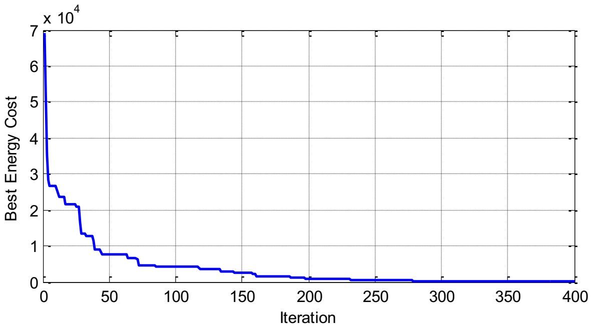

Fig. 6. MaxIt= 400; nPop = 200; a =0.7

Algorithm I : Artificial Bee Colony Scheduling Algorithm

```
% Inputs
LMin;  LMax;  VMList [ ]; HostList[ ]
 % Artificial Bees Colony Parameters
MaxIt     % Maximum Number of Iterations
nPop      % Population Size ;  a     % Acceleration Coefficient
% Initialization
% Generate Initial Population
For i=1 to nPop
   pop(i).position=PositionFunction(LMin, Lmax)
   pop(i).cost=CostFunction(pop(i).Position)
   If pop(i).Cost<=BestSol.Cost
      BestSol=pop(i)
   End
End
% Initialize abandonment counter
% Artificial Bees Colony Iterations
For it=1 to MaxIt
   For i=1to nPop
      Choose k randomly, not equal to i ;   Define Acceleration coefficient
      Choose new bee Position ;     Evaluate new solution found
   End
  Calculate fitness F and selection likelihoods P
   For i=1to nPop
       $F(i) = 1/(1 + f(x_i))$
   End
  P=F/sum(F)
  % Look for new onlooker Bees
   For m=1to onlooker
      Select Source Site;   Choose k randomly, not equal to i
      Select new Bee Position ; Evaluate new bee cost; keep better bee
    End
    % Update Scout Bees ;   % Update Best Solution  ;  % Store Best Cost
End
```

V. Conclusion

In summary, this study enabled us to evaluate the evolution of energy efficiency within the Cloud environment and the definition of a virtual machine scheduling algorithm in order to provide a better quality of service Thanks to a better processing time. Moreover, several observations and conclusions are still not exploited. We plan in the near future to implement all the findings established during the evaluation phase in order to come up with a global and efficient solution both to the energetic aspect and on the responsiveness aspect.

Acknowledgment

The authors would like to express their thanks to the anonymous referees for their valuable suggestions.

References

[1] A. T. Velte, T. J. Velte, and R. C. Elsenpeter, *Cloud computing a practical approach*. New York: McGraw-Hill, 2010.

[2] D. C. Marinescu, "Cloud Computing: Applications and Paradigms," in *Cloud Computing*, Elsevier, 2013, pp. 99–130.

[3] G. Quan and X. Hu, "Energy efficient fixed-priority scheduling for real-time systems on variable voltage processors," in *Proceedings of the 38th annual Design Automation Conference*, 2001, pp. 828–833.

[4] D. Kliazovich, S. T. Arzo, F. Granelli, P. Bouvry, and S. U. Khan, "e-STAB: Energy-Efficient Scheduling for Cloud Computing Applications with Traffic Load Balancing," 2013, pp. 7–13.

[5] J. Subirats and J. Guitart, "Assessing and forecasting energy efficiency on Cloud computing platforms," *Future Gener. Comput. Syst.*, vol. 45, pp. 70–94, Apr. 2015.

[6] I. S. Moreno and J. Xu, "Neural Network-Based Overallocation for Improved Energy-Efficiency in Real-Time Cloud Environments," 2012, pp. 119–126.

[7] R. Karthikeyan and P. Chitra, "Novel heuristics energy efficiency approach for cloud Data Center," in *Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on*, 2012, pp. 202–207.

[8] A. Horri, M. S. Mozafari, and G. Dastghaibyfard, "Novel resource allocation algorithms to performance and energy efficiency in cloud computing," *J. Supercomput.*, vol. 69, no. 3, pp. 1445–1461, Sep. 2014.

[9] M. Guzek, D. Kliazovich, and P. Bouvry, "A holistic model for resource representation in virtualized cloud computing data centers," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, 2013, vol. 1, pp. 590–598.

[10] Y. Yin, W.-H. Wu, T. C. E. Cheng, C.-C. Wu, and W.-H. Wu, "A honey-bees optimization algorithm for a two-agent single-machine scheduling problem with ready times," *Appl. Math. Model.*, vol. 39, no. 9, pp. 2587–2601, May 2015.

[11] W. Tian, Q. Xiong, and J. Cao, "An online parallel scheduling method with application to energy-efficiency in cloud computing," *J. Supercomput.*, vol. 66, no. 3, pp. 1773–1790, Dec. 2013.

[12] M. Fesanghary, S. Asadi, and Z. W. Geem, "Design of low-emission and energy-efficient residential buildings using a multi-objective optimization algorithm," *Build. Environ.*, vol. 49, pp. 245–250, Mar. 2012.

[13] C. Diakaki, E. Grigoroudis, and D. Kolokotsa, "Towards a multi-objective optimization approach for improving energy efficiency in buildings," *Energy Build.*, vol. 40, no. 9, pp. 1747–1754, Jan. 2008.

[14] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct. Multidiscip. Optim.*, vol. 26, no. 6, pp. 369–395, Apr. 2004.

[15] M. Tian *et al.*, "Multi-objective optimization of injection molding process parameters in two stages for multiple quality characteristics and energy efficiency using Taguchi method and NSGA-II," *Int. J. Adv. Manuf. Technol.*, vol. 89, no. 1–4, pp. 241–254, Mar. 2017.

[16] V. Josyula, *Cloud computing: automating the virtualized data center*. Indianapolis, IN: Cisco Press, 2012.

[17] Z. Mahmood, Ed., *Cloud Computing*. Cham: Springer International Publishing, 2014.

[18] D. P. Vidyarthi, Ed., *Scheduling in distributed computing systems: analysis, design & models*. New York: Springer, 2009.

[19] C. Xu and F. C. Lau, *Load balancing in parallel computers: theory and practice*, vol. 381. Springer Science & Business Media, 1996.

[20] H. Kameda, J. Li, C. Kim, and Y. Zhang, *Optimal Load Balancing in Distributed Computer Systems*. London: Springer London, 1997.

[21] T. Bourke, *Server load balancing*, 1st ed. Beijing ; Sebastopol, Calif: O'Reilly, 2001.

[22] D. Kliazovich, P. Bouvry, and S. U. Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers," *J. Supercomput.*, vol. 62, no. 3, pp. 1263–1283, Dec. 2012.

[23] O. B. Haddad, A. Afshar, and M. A. Mariño, "Honey-Bees Mating Optimization (HBMO) Algorithm: A New Heuristic Approach for Water Resources Optimization," *Water Resour. Manag.*, vol. 20, no. 5, pp. 661–680, Oct. 2006.

[24] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications," *Artif. Intell. Rev.*, vol. 42, no. 1, pp. 21–57, Jun. 2014.

[25] A. Ragmani, A. El Omri, N. Abghour, K. Moussaid, and M. Rida, "A global performance analysis methodology: Case of cloud computing and logistics," in *3 rd International Conference (GOL), 2016*, Fes Morocco, 2016, pp. 1–8.

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

[27] M. Chinnici and A. Quintiliani, "An Example of Methodology to Assess Energy Efficiency Improvements in Datacenters," 2013, pp. 459–463.

[28] G. Taguchi, S. Chowdhury, Y. Wu, S. Taguchi, and H. Yano, *Taguchi's quality engineering handbook*. Hoboken, N.J. : Livonia, Mich: John Wiley & Sons ; ASI Consulting Group, 2005.

[29] A. Ragmani, A. El Omri, N. Abghour, K. Moussaid, and M. Rida, "A performed load balancing algorithm for public Cloud computing using ant colony optimization," in *2nd International Conference (CloudTech 2016)*, Marrakech, Morocco, 2016, pp. 221–228.