Energy Management Service Layer for Cloud Computing Costs Reduction

Viviane T. Nascimento*, Andre L. V. Gimenes*, Tereza Cristina M. B. Carvalho* and Catalin Meirosu†
*University of São Paulo, São Paulo, Brazil
Email: vianetn@larc.usp.br, gimenes@pea.usp.br, terezacarvalho@usp.br
†Ericsson Research, Stockholm, Sweden.
Email: catalin.meirosu@ericsson.com

***Abstract*—Reducing the energy-related costs is a significant concern for Cloud Computing providers. Although many efficiency techniques had been evaluated, incentives for demand management according to energy sector information is not usual. The energy prices and availability vary according to daytime and geographical position, which implies in different prices for the resources in distinct periods. The higher flexibility of the Cloud Computing capacity regarding energy provider enables the adaptation of the infrastructure to take advantage of the energy prices fluctuation and the different sources availability. This work defines an infrastructure layer that allows negotiation and management of computing resources based on energy information for various time intervals. The proposed solution sets processing plans and establishes scheduling conditions to seize the most efficient usage of energy. A model that translates the customer contract information, searches for the lowest energy prices, and allocates the services based on these searches, is presented to reinforce our approach. Our results point out financial savings of almost 20% in comparison to non-energy management techniques.**

I. Introduction

The ICT sector is one of the greatest energy consumers in the world, as studies estimate that the industry consumed almost 7.4% of all the generated energy in 2012; this amount is expected to reach 12% until 2017 [1]. Related to the Cloud Computing, the demand growth for its services, as more business migrates to this model, could rank the sector among the six largest energy consuming countries in the world[2]. The high operational costs associated with the infrastructure powering pushes the providers to seek strategies to reduce the power consumption [3].

The energy efficiency in infrastructures are typically addressed as energy saving techniques, Data Centers (DCs) efficient architecture and the usage of renewable energy sources [4]. However, despite the evaluation of many strategies for the efficient consumption of energy, no approach considers the energy as a manageable and contractable resource for Cloud Computing.

This work proposes the management of different contracts for a Cloud Computing environment based on energy costs reduction. We define an infrastructure layer, named Energy as a Service (EaaS), that applies the energy inputs, such as energy sources availability, energy prices, geographical infrastructure location and load control, to manage the contracts. As far as we are aware, no work proposes a new service layer to deal with the specificities of the energy consumption for Cloud Computing environments.

Adapted from the energy sector, the EaaS performs demand management of contracts to change the Cloud energy load. Through the definition of service levels and quantitative terms for the Cloud Computing resources provisioning, the EaaS establishes goals for the energy purchasing and contract allocation.

The adjustment of the infrastructure to specific demand periods, as lowest prices of the energy generated, terms of higher generation of renewable sources, and geographical advantages, enables to adapt the services allocation and to incentive the consumers to change their contract patterns according to the energy requirements. The elastic property of the Cloud Computing resources provisioning enables to allocate and release the computing infrastructure according to demand requirements [3].

This paper is structured as follows: Section II describes the Demand Side Management and some energy approaches for DCs. Section III evaluates the energy as a separated manageable resource and the concept of the EaaS. Section IV evaluates the solution for the energy layer management and Section V shows the Use Case developed to the proposed solution. Section VI concludes the paper and presents the future steps for the work.

II. State of the Art

The National Institute of Standards and Technology (NIST) defines Cloud Computing as a model to enable continuous, convenient and on-demand access to a shared pool of computational resources that can be provisioned and released with minimal management effort or service provider interaction [5]. The same institute defines that standard mechanisms must access the resources and assigned dynamically according to the consumers demand. The services must provision and releasing must be easy, and the resources usage offered in a countable way [5].

The resources must be, therefore, provided in a way that enables the consumer to measure its quality. Jennings and Stadler [3] cite the resources management as one of the most challenging tasks in a cloud computing environment. The same work cites computing, networking, storage as manageable cloud computing related infrastructure; focusing on energy

978-1-5090-5137-3/16/$31.00

efficiency of the structure, the energy is also related as a manageable resource of the cloud.

The characteristics of cloud computing services provisioning, especially elasticity, and the constant measurement and control of the resources [5] are opportunities to manage the power load and the state of infrastructure to take advantage of demand management side programs. The Demand Side Management (DSM) is a behavioral incentive program focused on efficient use of resources and influences changes in the consumer energy usage to produce shifts in the time and load of the utility [6].

The DSM deployment is dependable of consumers response to different incentives. The Demand Response (DR) programs are one way to deploy the management of the energy load by final consumers, and they are grouped into Incentive-Based Programs and Price-Based Programs [7]. The Incentive-Based Programs encourage load modification by participants to reduce peak load or in situations that jeopardize energy providing by the utility [7]. The Price-Based programs promote the load change pattern by different energy prices during established time interval [7].

Through the participation in DR programs, consumers can reduce his energy consumption based on load reduction strategies and change the time-period of energy consumption [8]. The on-demand approach of the Cloud Computing environment enables to adapt the services allocation based on different energy prices and availability. Considering this perspective, the demanded contracts can be offered to improve energy usage and reduce its related costs.

The variation of prices and availability of energy sources are cited as incentives to services allocation. Lucanin and Brandic [9] propose a Cloud Computing aware of the different energy prices, and that allocates its services according to the electricity availability. The work defines the physical allocation, named green instance, that is scheduled according to the resources and energy usage. The proposal evaluates the scheduling of services in a cloud computing by the electricity variation during time intervals. However, the approach lacks the maturity to deal with the scheduling techniques to address different energy sources and geographically spread DCs. The same approach does not detail how to handle the various contract profiles required to a cloud computing environment.

Different works cite the schedule of resources usage based on energy availability and prices variation [10], [12]. The scheduling of jobs for DCs, to adjust the resources allocation to energy availability, is one approach for the intermittency of renewable sources [10]. The work presents a scheduling algorithm as a solution for the renewable sources deployment, defining time intervals for allocation based on the availability of these sources. Despite the approach, the work does not take advantage of the renewable resources intermittent behavior to provide different prices and time intervals for customers. Also, the work does not consider the high costs to purchase these type of source for users.

Lucanin and Brandic [13] address the DCs geographical approach as an advantage of the different prices for DCs allocation. The work presents the advantage to allocate the DCs in various regions to deploy different prices for the energy. The deployment of geographical locations is an advantage for the energy costs. Still, the work does not consider the consumption of renewable energy resources as another differential for the DCs powering and costs reduction.

The costs to purchase the resources and power the DC enables to determine the physical and time allocation for the contracted services. The dynamicity of the prices of the geographical and generation prices of the energy during a period influences the decision of the cloud controller. The allocation based on the commercial availability of the user helps to determine the type of service provided and the physical allocation[14], [15]. The dynamic pricing of the energy during specified intervals enables to predict the costs and to determine the power consumption deployment as a manageable resource in a DC.

III. Energy as a Service (EaaS)

The present work defines an infrastructure layer responsible for the Cloud Computing energy management, named EaaS. The EaaS deploys the infrastructure required and the energy information provided by the energy sector to take the best decisions of power consumption. Although the energy service layer is defined as a separated layer, it exchanges information with the other layers to gather information about the required infrastructure to run the services.

The information shared is based on the knowledge of the capacity provisioning and the constant monitoring of the computing resources. The interaction between these two resources are not restricted to the computational infrastructure; the information exchanged determines how to power the entire infrastructure including air cooling, heating, etc. Thus, the information traded enables to predict energy overload situations and energy efficiency opportunities.

The computing resource and the related structure demanded to process the services enable to predict the resources provisioning to fulfill the contracts requirements. The capacity predicted to run the services determine management strategies, including the amount of power required to maintain the infrastructure, energy consumption efficiency and costs minimization. The constant monitoring of the structure also allows to avoid energy surplus and sustain the infrastructure in safe consumption levels.

The customers define contracts terms that enable to predict the infrastructure provision, including how to allocate the services and quality terms agreed for the services provisioning. Service levels are determined to guarantee the quality of the service provided. The Service Level Agreement (SLA) is negotiated regarding the time interval for the services allocation, latency -referring the maximum allocation range -, services priority, among others.

The EaaS manages the contracts quality conditions to guarantee the energy provisioning of the infrastructure. Therefore, EaaS takes advantage of the knowledge of the capacity deployment to turn the energy-related terms negotiable for the Cloud

Computing environment, such as the type of energy source (renewable or non-renewable) and energy efficiency levels.

The information trade among the infrastructure layers and the contract terms determine specificities for the energy management. The energy service layer deploys the capacity and contracts data to establish the amount of power to run the whole infrastructure, its relate quality specificities, and negotiate the powering capacity requirements with the energy sector.

Considering the energy sector provided information, such as sources availability, the amount of power generated and prices for the consumers, EaaS sets the amount of energy to be purchased, the sources demanded, time requirements and the quality of energy. The constant negotiation with the energy market establishes the time and price requirements, and the geographical allocation for the services allocation. The interface with the energy market also enables to foresee lowest prices and greater renewable energy generation, and rapidly allocate the services in favorable conditions.

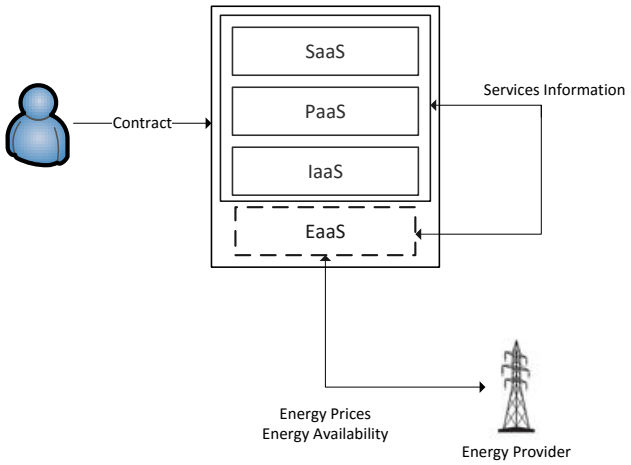


Fig. 1. Energy as a Service layer definition

Figure 1 shows the inclusion of the EaaS layer for the Cloud Computing. The contract inputs provide the resources to be provisioned by the cloud, including energy requirements. Although the energy management is an independent service layer, it is connected to the other layer to exchange information about the computing infrastructure required. The shared information represents the inputs for the EaaS layer to take decisions related to the energy consumption of the Cloud infrastructure. The information exchange between the EaaS and other layers represents the Services Information description in the figure.

EaaS interacts with the energy sector to receive information required to decide how to power the infrastructure. The two parts exchange information about the amount of energy to be purchased, availability, prices, quality and the grid situation. This interface enables to negotiate the best usage terms for the energy resources, including the advantage to deal with different power utilities

IV. Energy Efficiency Cloud Workflow

To manage energy in a Cloud Computing environment, a scheduling workflow, based on the balance of energy availability and prices, contract terms and usage of resources in DCs, is proposed. The workflow considers a system with time epochs t=0,1,2,...,n, each t representing the scheduling interval for C contracts. The contracts must be allocated on N geographically distributed DCs.

The first stage is the receipt of contracted information and its translation into energy terms. The contracts specify computing resources and requirements related to energy usage. The management system translates this information onto job descriptors, composed of power and computational resources required, time interval, latency, energy efficiency level and energy sources.

The solution establishes three allocation plans to deal with the different time and prices requirements contracted terms by users. Named reserved, flexible and on-demand, the plans set different prices level, according to processing time and energy availability. The plans do not define a priority for jobs allocation but indicate the willingness to take advantage of the lowest energy prices.

The reserved plan previously provides the period for service allocation and enables the cloud services provider to estimate energy and computing resources deployment. The on-demand plan requires immediate distribution and does not consider the advantages of energy prices and the best periods for the allocation. The flexible plan enables the assignment during the time interval with the lowest energy price, easy availability, or the lower computational resources demand of the DCs.

Time requirements for the energy prices and purchasing define the jobs management terms. The electricity sector yields energy prices ($p(t)$) for S sources in different $\mathbf{t}$ scheduling epochs. Since the plans require different ways to manage the resources, its choice reflects on their allocation costs.

Service levels (Sl) define different costs grades, established by the energy availability, prices and type of source demanded. The cheapest Sl is set for the flexible plan, as an incentive to users contracting this plan; therefore, the on-demand plan has the higher Sl rate. The Sl for the reserved plan varies during the managed periods, and is determined by $p(t)$ of the contracted epoch.

Equation 1 estimates the users costs to process each contracted job. The type of contract plan (Sl) determines the allocation cost, the cost to free the structure required to run the job (Sc), the quantity of power needed to execute the job ($JobEnergy$) and energy price $p(t)$, provided by the electricity market. The allocation costs vary in line with the time. The total JobCost is calculated for the k time epochs, and T is the final period for the job allocation.

$$JobCost(t) = \sum_{t=ti}^{t=tf} Sl(t) * Sc + p(t) * JobEnergy \quad (1)$$

The decision of which DC allocates the job is made based on a comparison between the costs to process it and an average

cost to allocate the job on the DC. This comparison enables that the jobs are always assigned on the cheapest DC; this decision is made based on profits maximization. The prices comparison allows the jobs geographical assignments since the DCs can be assigned in different areas, take advantage of the renewable sources offer and reduce jobs migration costs.

Equation 2 determines the average cost for a job processing on the DCs (*EC*), deployed as comparison criteria for the allocation, pausing and migration of jobs. *EC* is determined by the time *t* and the DC (*N*). The amount of energy expenditure for *s* servers in idle state (*Pidle*) is the minimum cost for the DC. During jobs processing (*k*), its quantity *Job* in the *t* specified scheduling epoch, the price payed (*pe*) and the amount of energy predicted for the jobs processing (*JobEnergy*) provides the cost of the DC.

$$EC(t) = \frac{\sum_{s=1}^{S} Pidle * PriceDC + JobEnergy(t) * pe(t)}{Job} \quad (2)$$

To accompany the usage of resources by the DCs, EaaS receives their current status periodically. The energy expenditure status provides the information about current load demanded by the computational resources. The received status also allows maintaining the energy load in a security margin.

In exceptional cases of spare energy or computational resources during the day, flexible plans can be allocated immediately, avoiding the waste of purchased resources. Other cases, as the overload of the infrastructure or excessive on-demand contracting, the flexible plans can be paused and reprocessed in a more affordable time or migrated for less demanded DCs.

The sequence for contracts scheduling is shown in figure 2. The first three modules correspond to the contracts receive, translation of the terms and the description into jobs. The Energy Estimate Module translates the contracts into information related to the energy usage, based on the computing resources input by the user. By the Contract Metrics module, the contracts are defined as jobs.

The Jobs Management module organizes the jobs by processing plan and time requirements, and points out which jobs must be processed, according to the scheduling epochs. Also, the module controls paused jobs and regulate the energy to be allocated. This module provides jobs information for the Jobs Scheduler.

The Jobs Scheduler is the module that directs where to allocate jobs, interacts with the energy sector and the DCs. The module receives the jobs to be processed by the management modules and determines the allocation costs for each one, based on energy provided data. The module decides which DC is going to allocate the job focusing on the energy best usage, but does not establish how the DC executes the job.

V. Use Case

The goal of the model is to show how energy prices and costs determine the allocation of cloud computing contracts. For this model, we defined *C*=100 contracts randomly-

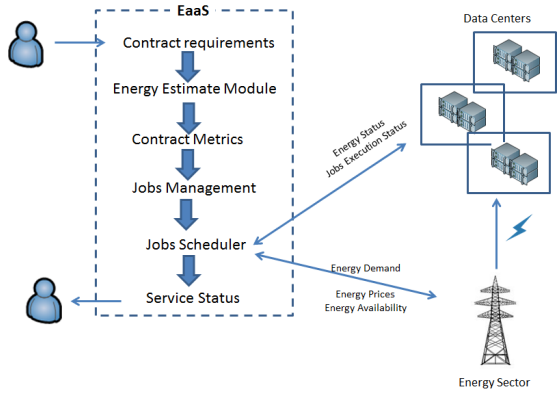


Fig. 2. Proposed Schema for the EaaS evaluation

generated, each one establishing cloud resources level, initial (*ti*) and final (*tf*) allocation hours, latency (*T*), type of plan and energy source required. The contracts state four different levels for the cloud resources: light, medium, high and highest and three options for the energy sources, renewable, non-renewable or indifferent.

Based on the four resources levels, the amount of energy to be deployed is defined. We set the energy consumption levels according to a typical high-performance server in 2016 [16]: *Pidle*=300 W and 1000 W at maximum load. The light resources contracting is estimated in *JobEnergy=Pidle+30%*Pidle*, medium in *JobEnergy=Pidle+75%*Pidle*, high level in *JobEnergy=Pidle+150%*Pidle* and the highest contracted resources consumes the maximum load.

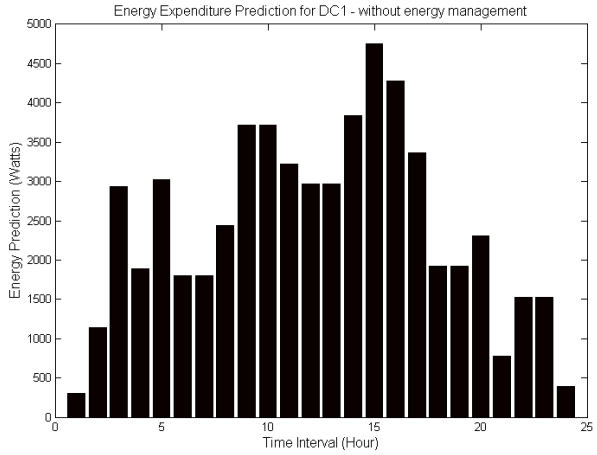
In terms of time restraints, reserved and on-demand plans define the initial and final processing hours, and flexible plans set the latency. The most appropriated interval for the flexible services processing is determined by equation 1, which defines the cheapest time for their allocation. To ease the management of the jobs, they are organized into queues. The system manages the queues according to their contracted plan and time requirements.

Based on Time of Use (ToU) demand response programs, energy prices were defined hourly (0<t<24)and *S*=2 sources, without specifying which source. The model predicts *N*=3 DCs: DC1 is powered only by renewable energy, DC2 powered only by non-renewable sources and DC3 that searches for the lower price, between the renewable and non-renewable prices established. Since the model does not consider jobs migration, DC1 allocates all the contracts that demand renewable energy sources. The other two types are assigned based on the comparison of costs for the processing interval of DC2 and DC3.

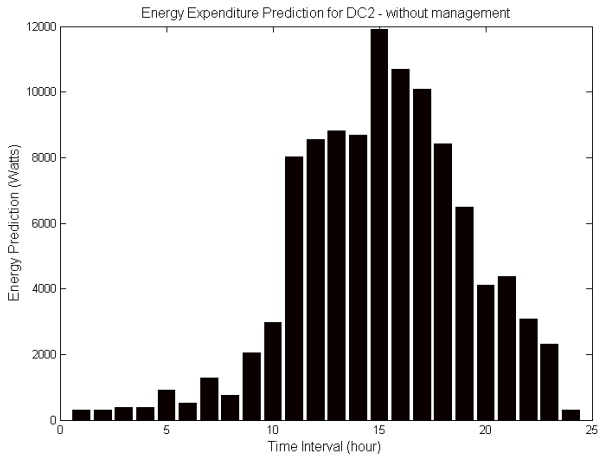
The model accounted 56 reserved contracts, 13 on-demand, and 31 flexible ones. Equation 1 is applied to determine the cost to process the jobs according to time requirements. The model calculates *Sl* as the difference between the lowest energy price and the price during the processing time. For on-demand jobs, *Sl*=1 and for flexible ones, *Sl*=0.1. Service Costs

(*Sc*) values ten currency units for all the contracts, independent of required plan.

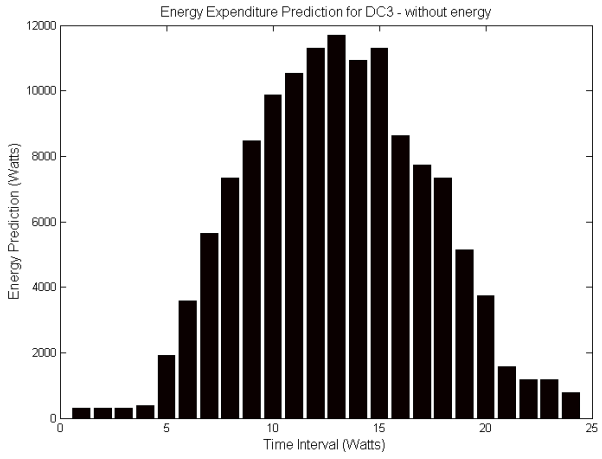
Figures 3 and 4 shows jobs distribution during the 24-hour processing interval. For figure 3, jobs are allocated on the first-



(a) Data Center 1

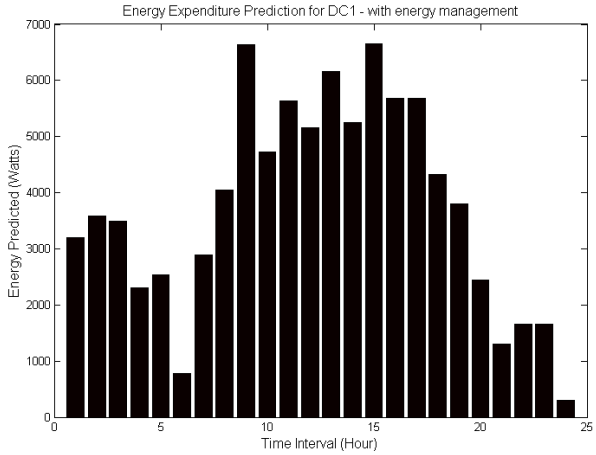


(b) Data Center 2

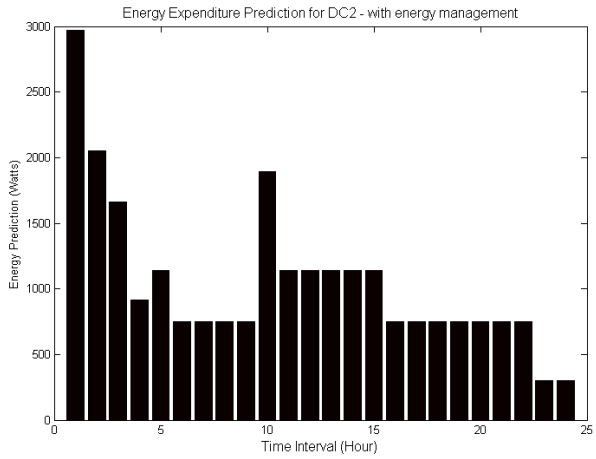


(c) Data Center 3

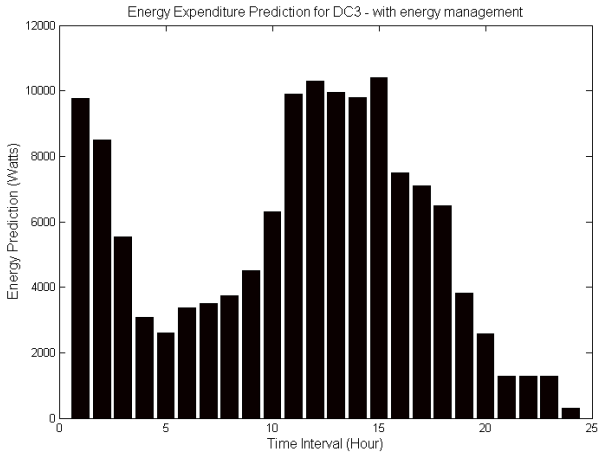
Fig. 3. Power Distribution per DC distributed in 24 hour interval without management



(a) Data Center 1



(b) Data Center 2



(c) Data Center 3

Fig. 4. Energy Consumption per DC distributed in 24 hour interval using energy management

TABLE I
COSTS PER DC AND TOTAL COST WITH AND WITHOUT ENERGY MANAGEMENT

Energy Costs($)	DC1	DC2	DC3	Total Cost
Without energy management	1.11	2.33	2.36	5.80
With energy management	1.69	0.52	2.44	4.65

come-first-served basis, without applying any management or searching for the lowest energy prices.

Figure 4 shows that the energy is managed to avoid the peak usage during the most energy-costly hours, such as afternoon, for non-renewable and dawn for renewable ones. The peak usage happens for DC3, the cheapest one, during the beginning of the day. The energy costs distribute the allocation during the dawn and morning, hours with the lowest prices, without compromising the jobs that demanded time intervals on their contracting process.

The full cost of both jobs allocation was determined by equation 2 that calculates the costs for each range and the sum of the values provided total cost for the 24-hour interval. For the non-energy management, the 24-hour cost is equal to 5.79 monetary units, and for the energy management method, the value is equal to 4.65 monetary units. Comparing both the costs, the management based on the energy prices presented an economy of almost 20%. Table I contains the energy-related costs for each DC and the total costs to maintain the 24-hour period of the Cloud Computing referred environment, with and without energy management.

VI. CONCLUSION AND FUTURE WORKS

This work presented the definition of an independent layer for the energy management of the Cloud Computing environment. The EaaS enables to negotiate the energy deployment regarding the lowest prices and greater sources availability. Also, the new layer allows the users to define energy terms by contract, which enables them to indicate the most appropriate approach for their business continuity.

The solution proposed for the EaaS implementation is the schedule of jobs, balancing costs and prices information related to the energy resources. Thus, the EaaS becomes an interface with the DCs, users and energy sector to search for the best opportunities to reduce the costs related to the energy usage. The scheduling method enables to take advantage of geographical spread DCs and the time intervals with the significant availability of renewable sources.

The model defined to demonstrate our approach shows the results for the management of contracts based on the more affordable energy prices. The savings, about 20% of financial savings, indicates significant gains for the cloud computing operator. The model enables to predict the amount of energy to be deployed on the environment and to balance the consumption in different DCs as well.

Our future steps include extending the model defined for the EaaS, increasing the gains opportunities. A method to enable the negotiation of the energy deployment with the final user and the future inclusion of the jobs migration, balancing costs and savings, from different DCs are the further steps of our work.

REFERENCES

[1] K. R. B. J. D. Pomerantz, G. Cook, "Clicking Clean: A Guide to Building the Green Internet," Greenpeace, Tech. Rep., 2015. [Online]. Available: http://www.greenpeace.org/usa/wp-content/uploads/legacy/Global/usa/planet3/PDFs/2015ClickingClean.pdf

[2] G. Cook, T. Dowdall, D. Pomerantz, and Y. Wang, "Clicking Clean: How Companies are Creating the Green Internet," Greenpeace, Tech. Rep., 2014. [Online]. Available: http://www.greenpeace.org/usa/wp-content/uploads/legacy/Global/usa/planet3/PDFs/clickingclean.pdf

[3] B. Jennings and R. Stadler, "Resource management in clouds: Survey and research challenges," *Journal of Network and Systems Management*, vol. 23, no. 3, pp. 567–619, 2015.

[4] A. Hammadi and L. Mhamdi, "A survey on architectures and energy efficiency in data center networks," *Computer Communications*, vol. 40, pp. 1–21, 2014.

[5] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST, Tech. Rep., 2011. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.800-145

[6] D. Goswami and F. Kreith, *Energy Efficiency and Renewable Energy Handbook, Second Edition*, ser. Mechanical and Aerospace Engineering Series. CRC Press, 2015. [Online]. Available: https://books.google.com.br/books?id=GtaYCgAAQBAJ

[7] R. Deng, Z. Yang, M. Y. Chow, and J. Chen, "A survey on demand response in smart grids: Mathematical models and approaches," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 570–582, June 2015.

[8] P. Siano, "Demand response and smart grids—a survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014.

[9] D. Lucanin and I. Brandic, "Take a break: cloud scheduling optimized for real-time electricity pricing," in *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, 2013, pp. 113–120.

[10] H. T. Minh and M. Samejima, "An evaluation of job scheduling based on distributed energy generation in decentralized data centers," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1172–1177.

[11] Í. Goiri, K. Le, M. E. Haque, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenslot: scheduling energy consumption in green datacenters," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 20.

[12] C. Joe-Wong, I. Kamitsos, and S. Ha, "Interdatacenter job routing and scheduling with variable costs and deadlines," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2669–2680, 2015.

[13] D. Lučanin and I. Brandic, "Pervasive cloud controller for geotemporal inputs," *IEEE Transactions on Cloud Computing*, vol. 4, no. 2, pp. 180–195, April 2016.

[14] S. Zaman and D. Grosu, "A combinatorial auction-based mechanism for dynamic vm provisioning and allocation in clouds," *IEEE Transactions on Cloud Computing*, vol. 1, no. 2, pp. 129–141, 2013.

[15] L. Mashayekhy, M. M. Nejad, D. Grosu, and A. V. Vasilakos, "An online mechanism for resource allocation and pricing in clouds," *IEEE Transactions on Computers*, vol. 65, no. 4, pp. 1172–1184, 2016.

[16] Supermicro, "Superserver 8028b-tr4f," Super Micro Computer, Inc., Tech. Rep., June 2016. [Online]. Available: http://www.supermicro.com/products/system/2U/8028/SYS-8028B-TR4F.cfm