Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 1, No. 1, January 2021

AN APPROACH TO BALANCE MAINTENANCE COSTS AND ELECTRICITY CONSUMPTION IN CLOUD DATA CENTERS

[1]B. Harish Kumar Reddy, [2]S. Somasekhar
[12]Assistant Professor Department Of CSE Dr. K. V. Subba Reddy Institute Of Technology, Kurnool.

ABSTRACT

We target the problem of managing the power states of the servers in a Cloud Data Center (CDC) to jointly minimize the electricity consumption and the maintenance costs derived from the variation of power (and consequently of temperature) on the servers' CPU. More in detail, we consider a set of virtual machines (VMs) and their requirements in terms of CPU and memory across a set of Time Slot (TSs). We then model the consumed electricity by taking into account the VMs processing costs on the servers, the costs for transferring data between the VMs, and the costs for migrating the VMs across the servers. In addition, we employ a material-based fatigue model to compute the maintenance costs needed to repair the CPU, as a consequence of the variation over time of the server power states. After detailing the problem formulation, we design an original algorithm, called Maintenance and Electricity Costs Data Center (MECDC), to solve it. Our results, obtained over several representative scenarios from a real CDC, show that MECDC largely outperforms two reference algorithms, which instead either target the load balancing or the energy consumption of the servers.

I. INTRODUCTION

The number and scale of data centers (DCs) are rapidly increasing, as data centers are the primary means through which companies can satisfy their increasing demand for computing and storage resources, either by building private data centers or by offloading applications and services to external cloud providers. A major issue is that data centers consume a large amount of energy and that consumption is expected to increase at a significant rate. It is estimated that data center electricity consumption will increase to roughly 140 billion kilowatt-hours annually by 2020, corresponding to about 50 large power plants, with annual carbon emissions of nearly 150 million metric tons. The financial impact for DC management is also huge, since a DC spends between 30% and 50% of its operational expenditure in electricity: the expected figure for the sector in 2020 is $13 billion per year of electricity bills (updated information can be found on the web portal of the U.S. National Resources Defense Council , The efficient utilization of resources in the data centers is therefore essential to reduce costs, energy consumption, carbon emissions and also to ensure that the quality of service experienced by users is adequate and adherent to the stipulated service level agreements. Efficiency is essential not only in single data centers, but also in geographically-distributed data centers, whose adoption is rapidly increasing. Major cloud service providers, such as Amazon, Google and Microsoft, are deploying distributed data centers to match the increasing demand for resilient and low-latency

cloud services, or to interconnect heterogeneous data centers owned by different companies, in the so-called "inter-cloud" scenario.

In this scenario, the dynamic allocation and migration of workload among data centers can help to reduce costs, moving the workload where the energy is less expensive/cleaner and/or cooling costs are lower: the cloud provider has the option of choosing the destination site based on different criteria upon the reception of the user request. Specifically, the increasing adoption of renewable energy plants is a great opportunity for a more efficient management of distributed data centers. Each data center can get its electricity from different electricity providers or can adopt on-site renewable energy sources, which provide green energy such as solar and wind [1].

Moving applications and services to data centers that are equipped with renewable energy sources (RES) can lead to several benefits both for the data center provider, which can reduce the costs of acquiring grid energy, and for the society in general, thanks to a more intense exploitation of green energy and the reduction of carbon emissions.

Unfortunately, workload assignment and migration in a distributed environment involve very complex decision processes due the time-variability of electricity cost, the workload variability both within single sites and across the whole infrastructure and, when the adoption of RES is possible, the intermittent nature of green energy generation. In a geographically-distributed scenario, the peak hours of renewable energy generation can be different in each data center, due to the variability of meteorological conditions and the different time zones. This generates the need for moving the workload to the sites where and when the green energy is currently available. Indeed, while energy storage units can be appropriately used to defer the utilization of green energy for some time, this postponement comes at a significant cost related to the charging and discharging of batteries.

The immediate usage of green energy is the most convenient option, but can be exploited only if the infrastructure is able to support the workload redistribution and if efficient algorithms are designed to drive these workload shifts. In the paper, we prove that it is possible to effectively use green energy to reduce operational cost with a smart workload distribution. In this paper, we present a novel approach for the efficient exploitation of RES in a distributed data center scenario, by adapting and refining a workload management strategy already proposed in the literature, i.e., EcoMultiCloud [2].

EcoMultiCloud includes a hierarchical architecture for the management of geographically-distributed data centers and a set of algorithms that drive the assignment and migrations of virtual machines (VMs) on the basis of the technical and business objectives defined by the management.

EcoMultiCloud is composed of two layers: at the lower layer, each site adopts its own strategy to distribute and consolidate the workload internally. At the upper layer, a set of algorithms—shared by all the sites—are used to evaluate the behavior of single sites and distribute the workload among them. This architecture offers several benefits, among which:

(i) scalability, because the bigger problem of workload allocation is decomposed into smaller intra-data center an inter-data center problems;

(ii) autonomy of single data centers, since each data center can adopt its own algorithm for internal allocation;

(iii) flexibility, since the algorithms can be easily adapted in accordance with the desired objectives.

In [2], the objectives that drove the workload redistribution were the load balancing among the sites and the minimization of costs; in that paper, the flexibility and feasibility of EcoMultiCloud were also verified with the support of analytical models. However, the availability of RES was not considered. In this paper, thanks to the flexibility characteristics mentioned above, we adapt
the EcoMultiCloud algorithm to exploit the presence of renewable energy plants specifically. The EcoMultiCloud algorithm proves to be suited to the new, more challenging scenario, in which RES is introduced. EcoMultiCloud accomplishes the objective to use renewable energy when and where it is generated; energy produced in excess with respect to what is consumed can be stored for future use. We analyzed a scenario with four data centers, located in various geographical areas and with different patterns of renewable energy generation. Performance results mainly concern the reduction in the use of grid energy and the cost savings that derive from this reduction. Results were derived when varying the size of photovoltaic panels and the size of batteries, which allowed us to determine the appropriate values of both parameters. Moreover, we evaluated the advantages deriving from the use of the novel strategy with respect to a strategy that does not exploit migrations and to a strategy that moves the workload randomly from the most costly data center to the other data centers. The main contribution of the paper is two-fold. On the one hand, the paper shows that, while it is intermittent and highly variable, renewable energy can be effectively used in geographically-distributed data centers to reduce operational costs due to the infrastructure power supply, given that a smart and dynamic workload management is adopted. On the other hand, the paper proves that EcoMultiCloud, used in the past only in scenarios with traditional power supply, is suited for the purpose and can be easily adapted to consider green energy in the scenario.

II. EXISTING SYSTEM

The dynamic allocation and migration of workload among data centers can help to reduce costs, moving the workload where the energy is less expensive/cleaner and/or cooling costs are lower: the cloud provider has the option of choosing the destination site based on different criteria upon the reception of the user request. Specifically, the increasing adoption of renewable energy plants is a great opportunity for a more efficient management of distributed data centers. Each data center can get its electricity from different electricity providers or can adopt on-site renewable energy sources, which provide green energy such as solar and wind [1]. Moving applications and services to data centers that are equipped with renewable energy sources (RES) can lead to several benefits both for the data center provider, which can reduce the costs of acquiring grid energy, and for the society in general, thanks to a more intense exploitation of green energy and the reduction of carbon emissions. Unfortunately, workload assignment and migration in a distributed environment involve very complex decision processes due the time-variability of electricity cost, the workload variability both within single sites and across the whole infrastructure and, when the adoption of RES is possible, the intermittent nature of green energy generation. In a geographically-distributed scenario, the peak hours of renewable energy

generation can be different in each data center, due to the variability of meteorological conditions and the different time zones. This generates the need for moving the workload to the sites where and when the green energy is currently available. Indeed, while energy storage units can be appropriately used to defer the utilization of green energy for some time, this postponement comes at a significant cost related to the charging and discharging of batteries.

The immediate usage of green energy is the most convenient option, but can be exploited only if the infrastructure is able to support the workload redistribution and if efficient algorithms are designed to drive these workload shifts. In the paper, we prove that it is possible to effectively use green energy to reduce operational cost with a smart workload distribution.

Disadvantages :

Load balancing schemes depending on whether the system dynamics are important can be either static and dynamic . Static schemes do not use the system information and are less complex.

III. PROPOSED SYSTEM

we present a novel approach for the efficient exploitation of RES in a distributed data center scenario, by adapting and refining a workload management strategy already proposed in the literature, i.e., EcoMultiCloud [2]. EcoMultiCloud includes a hierarchical architecture for the management of geographically-distributed data centers and a set of algorithms that drive the assignment and migrations of virtual machines (VMs) on the basis of the technical and business objectives defined by the management. EcoMultiCloud is composed of two layers: at the lower layer, each site adopts its own strategy to distribute and consolidate the workload internally. At the upper layer, a set of algorithms—shared by all the sites—are used to evaluate the behavior of single sites and distribute the workload among them. This architecture offers several benefits, among which:

(i) scalability, because the bigger problem of workload allocation is decomposed into smaller intra-data center an inter-data center problems;
(ii) autonomy of single data centers, since each data center can adopt its own algorithm for internal allocation;
(iii) flexibility, since the algorithms can be easily adapted in accordance with the desired objectives.

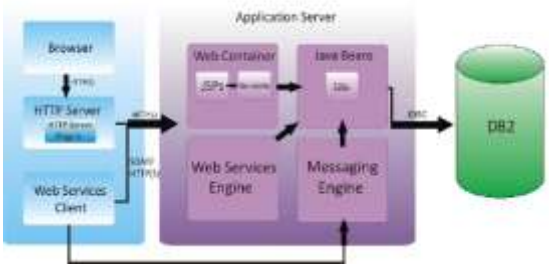
In [2], the objectives that drove the workload redistribution were the load balancing among the sites and the minimization of costs; in that paper, the flexibility and feasibility of Eco Multi Cloud were also verified with the support of analytical models. However, the availability of RES was not considered. In this paper, thanks to the flexibility characteristics mentioned above, we adapt the EcoMultiCloud algorithm to exploit the presence of renewable energy plants specifically. The EcoMultiCloud algorithm proves to be suited to the new, more challenging scenario, in which RES is . introduced. EcoMultiCloud accomplishes the objective to use renewable energy when and where it is generated; energy produced in excess with respect to what is consumed can be stored for future use.

We analyzed a scenario with four data centers, located in various geographical areas and with different patterns of renewable energy generation. Performance results mainly concern the reduction in the use of grid energy and the

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 1, No. 1, January 2021

cost savings that derive from this reduction. Results were derived when varying the size of photovoltaic panels and the size of batteries, which allowed us to determine the appropriate values of both parameters. Moreover, we evaluated the advantages deriving from the use of the novel strategy with respect to a strategy that does not exploit migrations and to a strategy that moves the workload randomly from the most costly data center to the other data centers. The main contribution of the paper is two-fold. On the one hand, the paper shows that, while it is intermittent and highly variable, renewable energy can be effectively used in geographically-distributed data centers to reduce operational costs due to the infrastructure power supply, given that a smart and dynamic workload management is adopted. On the other hand, the paper proves that Eco Multi Cloud, used in the past only in scenarios with traditional power supply, is suited for the purpose and can be easily adapted to consider green energy in the scenario.

IV. SYSTEM ARCHITECTURE



V. IMPLEMENTATION

This section is devoted to the description of the system architecture. As previously mentioned, the multi-site load management strategy considered in this paper is EcoMultiCloud [2]. EcoMultiCloud consists of a two-layer architecture in which the upper layer is used to exchange information among the different DCs so as to properly drive the distribution of VMs among the sites, while the lower layer is used to allocate the workload within single DCs. At the lower layer, the DCs can use any load management strategy. In this paper, the DCs use the decentralized/self-organizing approach presented in [14] for the consolidation of the workload in a DC. The single DC solution dynamically consolidates VMs to the minimum number of servers and allows the remaining servers to enter low consuming sleep modes. This approach has been proven to be very efficient for the energy consumption reduction of individual data centers. At the lower layer, key decisions regarding the local data center are delegated to the servers, which autonomously decide whether or not to accommodate a VM or trigger a VM migration. At the upper layer, global strategies are implemented by making decisions about workload allocation and inter-DC VM migrations. The decisions are made by combining some general information about single DCs. The hierarchical architecture, organized in two independent layers that exchange some (limited) information, allows upper layer algorithms to be modified independently on the lower layer algorithms. In this way, different strategies can be adopted locally at the DCs.

Conversely, improvements of single sites can be implemented without any explicit notification or involvement of the upper layer, i.e., of other DCs, provided that information exchange between layers is maintained. The reference scenario is depicted in Figure 1, which

shows the upper and lower layers for two interconnected DCs. At each DC, a data center manager (DCM) runs the algorithms of the upper layer. The DCM integrates the information coming from the lower layer and uses it to implement the functionalities of the upper layer. The DCM communicates with the local manager (LM) and acquires detailed knowledge about the current state of the local DC, for example regarding the usage of host resources and the state of running VMs. Then, the DCM extracts relevant high level information about the state of the DC and transmits this high level information to all the other DCMs (upper layer). The algorithms at the upper layer combine the collected information and make decisions about the distribution of the workload among the DCs. The assignment algorithm is used to decide to which DC a new VM should be assigned. Once the VM is delivered to the target site, the LM runs the lower layer algorithms to assign the VM to a specific host.

Main Modules:-

1. USER MODULE :

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

2. System Model :

There are several cloud computing categories with this work focused on a public cloud. A public cloud is based on the standard cloud computing model, with service provided by a service provider . A large public cloud will include many nodes and the nodes in different geographical locations. Cloud partitioning is used to manage this large cloud. A cloud partition is a subarea of the public cloud with divisions based on the geographic locations. with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition.

3. Main controller and balancers:

The load balance solution is done by the main controller and the balancers.

The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information. Since the main controller deals with information for each partition, smaller data sets will lead to the higher processing rates. The balancers in each partition gather the status information from every node and then choose the right strategy to distribute the jobs.

4. Cloud Partition Load Balancing Strategy:

When the cloud partition is idle, many computing resources are available and relatively few jobs are arriving. In this situation, this cloud partition has the ability to process jobs as quickly as possible so a simple load balancing method can be used. There are many simple load balance algorithm methods such as the Random algorithm, the Weight Round Robin, and the Dynamic Round Robin

The Round Robin algorithm is used here for its simplicity.

VI. CONCLUSION

We consider Eco Multi Cloud, a flexible load management tool that implements multi-objective load management strategies, and we adapt it to the presence of renewable energy sources to power data centers. Our work aims at achieving cost reduction in the load allocation process in a multi-data center scenario, where VMs are assigned to a given data center by considering both energy cost variations and the presence of local renewable energy production, in order to reduce the energy bill. Performance is investigated for a specific infrastructure consisting of four data centers. Results show that geographical data centers can significantly benefit from renewable energy sources when a smart load allocation strategy is implemented. Our results show that more than 60% cost can be saved even with a limited size of PV panels, with energy bill reduction as high as almost 100% under proper dimensioning of the PV panel and battery capacity. Scaling up to realistic scenarios, where data centers are made up of thousands of servers each, this means that the energy cost-saving can be of the order of several million dollars per year. Furthermore, combining the use of RE with a dynamic workload distribution by means of the application of migration policies allows further reduction of the energy bill by up to 17%, even without any storage and regardless of the type of adopted strategy. With large RE generators, the introduction of migration policies becomes less beneficial. Hence, the paper proves the feasibility of the introduction of renewable energy to reduce the operational costs of a complex infrastructure such as cloud data centers with geographically-distributed sites. Renewable energy can be effectively introduced given that a smart workload management is implemented; an approach like Eco Multi Cloud is suited for this purpose.

Our results also highlight how the selection of the DCs in which the PV panels are installed is critical and different cost savings can be obtained, depending on the variable energy prices, the local RE production level and the PUE of each DC. In addition, given the same initial investment for a RE generator with a given capacity, although it is not necessary to equip all the DCs with some PV panels, it is more convenient to distribute the installation of PV panels on a fraction of DCs, rather than consolidating the R E generation capacity on a single DC. Future work is required to investigate whether cost-saving might further benefit from the implementation of load management strategies that consider also the forecast future RE production to perform the VM assignment.

REFERENCES

[1] R. Hunter, The why of cloud, http://www.gartner.com/ DisplayDocument?doc cd=226469&ref= g noreg, 2012.

[2] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, Cloud computing: Distributed internet computing for IT and scientific research, Internet Computing, vol.13, no.5, pp.10-13, Sept.-Oct. 2009.

[3] P. Mell and T. Grance, The NIST definition of cloud computing, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, 2012.

[4] Microsoft Academic Research, Cloud computing,http://libra.msra.cn/Keyword/6051/cl

oud-computing?query=cloud%20computing, 2012.

[5] Google Trends, Cloud computing, http://www.google.com/trends/explore#q=cloud %20computing, 2012.

[6] N. G. Shivaratri, P. Krueger, and M. Singhal, Load distributing for locally distributed systems, Computer, vol. 25, no. 12, pp. 33-44, Dec. 1992.

[7] B. Adler, Load balancing in the cloud: Tools, tips and techniques, http://www.rightscale. com/info center/white-papers/Load-Balancing-in-the-Cloud.pdf, 2012