A Model for Energy-Awareness in Federated Cloud Computing Systems with Service-Level Agreements

Alessandro Ferreira Leite, Alba Cristina de Melo, Christine Eisenbeis, Claude Tadonki

A Model for Energy-Awareness in Federated Cloud Computing Systems with Service-Level Agreements

Alessandro Ferreira Leite[1,2]
Supervisors
Alba Cristina Magalhães Alves de Melo[1], Christine Eisenbeis[1,3], and Claude Tadonki[4,5]

[1] Université Paris-Sud 11
[2] University of Brasilia
[3] INRIA Saclay
[4] MINES ParisTech / CRI

Abstract. As data centers increase in size and computational capacity, numerous infrastructure issues become critical. Energy efficient is one of these issues because of the constantly increasing power consumption of CPUs, memory, and storage devices. A study shows that the whole energy consumed by data centers will be extremely high and it is like to overtake airlines in terms of carbon emissions. In that scenario, Cloud computing is gaining popularity since it can help companies to reduce costs and carbon footprint, usually distributing execution of services across distributed data centers. The research aims of this work are to propose and evaluate a Model for Federated Clouds that takes into account power consumption and Quality of Service (QoS) requirements. In our model, the energy reduction shall not result in negative impacts to the agreements between Cloud users and Cloud providers. Therefore, the model should ensure both energy-efficiency and QoS parameters, which sets up possibly conflicting objectives.

1 Introduction

Nowadays, the energy cost can be seen as one of the major concerns of data centers, since it is sometimes nonlinear with the capacity of those data centers, and it is also associated with a high amount of Carbon emission (CO_2). Some projections considering the data center energy-efficiency [1] show that the total amount of electricity consumed by data centers in the next years will be extremely high and that the associated carbon emissions would reach unprecedented levels.

Depending on the efficiency of the data center infrastructure, the number of watts that it requires can be from three to thirty times higher than the number of watts needed for computations [2]. And it has a high impact on the total operation costs [3], which can be over 60% of the peak load.

The question is how to increase the energy efficiency of the whole data centers without sacrificing Quality of Service (QoS) requirements, both for economical reasons and for making the IT environment sustainable [4]. Answering that questions is difficult since there are many variables that contribute to the power consumption of a resource. For instance, the power consumption of a resource does not depend only on its architecture or on the application it is running but also it depends on its position in the data center and on the temperature of the data center [5].

Most of the researches on energy efficiency have focused on hardware. The hardware energy efficiency has significantly improved, with particularly high gains in the energy efficiency of hardware power consumption. However, whereas hardware is physically responsible for power consumption, hardware operations are guided by software, which is indirectly responsible for energy consumption [6].

Moreover, many energy-efficient computing approaches focus on single-objective optimizations, without considering the QoS parameters. Energy-saving schemes that result in too much degradation of the system performance or in violations of Service-Level Agreements (SLA) parameters would eventually cause the users to move to another provider. Therefore, there is a need to reach a balance between the energy savings and the costs incurred by these savings in the execution of the applications.

In that context, Cloud computing is gaining popularity since it can help companies to reduce costs and carbon footprint, usually distributing the execution of their services across distributed data centers. A Cloud computing data centers usually employ virtualization techniques to provide computing resources as utilities to provision computational resources on-demand, and auto-scaling techniques to dynamically allocate resources to applications accordingly to the load, removing resources that would otherwise remain idle and wasting power consumption.

In order to support a large number of consumers or to decentralize management, Clouds can be combined, forming a Federated Cloud environment. A Federated Cloud can move services and tasks among Clouds in order to achieve its goals. These goals are usually described as QoS metrics, such as minimum execution time, minimum price, availability, minimum power consumption and minimum network latency, among others.

Federated Clouds are an elegant solution to avoid overprovisioning, thus reducing the operational costs in an average load situation, while still being able to give QoS guarantees to the users. In that case, our research aims are to propose and evaluate a model for Federated Clouds that takes into account power consumption and SLA constraints. In this proposal, the energy reduction shall not result in negative impacts to the Service-Level Agreements (SLA) between Cloud users and Cloud providers. Therefore, the model should ensure both energy-efficiency and QoS parameters, which sets up possibly conflicting objectives.

2 Related Works

Basically there are two approaches to reduce power consumption in a data center. The first one is the method of energy-aware hardware design, which can be carried out at various levels, such as device-level power reduction, circuit and logic level intelligent power management and architecture power reduction. The second approach is the method of power-aware software design, known as Dynamic Voltage and Frequency Scaling (DVFS) [7], including the Operating System, the applications and resource allocation in general.

In [8], a DVFS and temperature aware load balancing technique is presented to constrain core temperatures. The approach lets each core working at the maximum available frequency until a temperature threshold is reached. Experiments in a cluster with dedicated air conditioning unit show that a cooling saving of 57% can be achieved with 20% of timing penalty.

In [9], Mitrani proposes a dynamic operating policy using Queueing Theory that considers power consumption and users defect, if they have to wait too long before the service starts. The servers are switched to on/off in block and they are put in a reserve state accordingly with the workload of the system. Considering numerical experiments, the results show that the cost of losing a request increases with the workload and that the benefit of using a dynamic policy, compared to leaving all servers powered on, is larger for lighter loads than for heavier ones. If the number of servers is fixed to the maximum in lighter load scenario, the powered on reserves are not necessary. And, for heavier scenarios, the servers should be powered on as soon as a queue appears.

In [10], Khazaei, Misic and Misic model a Cloud data center as an $M/G/m/m+r$ queue system, proposing an analytical technique based on an approximate Markov chain model to evaluate aspects related to performance indicators without imposing restrictions to the number of servers and assuming a general service time for requests. The aim of the authors was to evaluate the probability distribution of the response time, the number of tasks in the system, and the size of the buffer needed for the blocking probability, using a combination of transformed base analytical model and a homogeneous and ergodic Markov chain. The results show that the impact of the buffer size becomes imperceptible as the number of server increases, the probability of blocking a task decreases when the buffer size increases, and that, considering SLA parameters, it is better to have distinct homogeneous Clouds instead of having one heterogenous Cloud.

In [11], a strategy based on Game Theory is proposed for resource allocation in Horizontal and Dynamic Federated Clouds. Two asynchronous utility games are proposed. The first one, called UtilMax$\underline{p}$CP game, aims to maximize the total profit for the buyer Cloud Provider (CP) and the second one (UtilMax$\underline{c}$CP game) tries to maximize the social welfare for the seller CPs. In that Federation environment, the CPs make decisions based on local knowledge and preferences and global decisions are achieved through interactions among them. The main goal is to seek an equilibrium for the whole system.

The results obtained with mathematical simulation show that the UtilMaxpCP game achieved the highest welfare, and the UtilMaxcCP game obtained the highest return on investment and it was more cost-effective.

3 Discussion and Research Approach

Recently, Cloud computing has been receiving a lot of attention since it is able to provide utility computing in an elastic environment. The advantages of Cloud computing can be obtained at zero cost since many of the Public Clouds provide free usage slots, allowing users to run their applications for free in Cloud environments. Also, many Clouds can be put together and seen as a unique environment. Public Clouds can be suitable for execute scientific applications as they provide virtually unlimited amount of computing resources on-demand and nearly realtime, especially for users whose peak computing resources needs exceed the capacity of available resources.

However, even though Public Clouds have numerous advantages, they also have several disadvantages for scientific applications. The first disadvantage is that there is significantly evidence that they cannot produce repeatable and reproducible scientific results [12]. The Cloud Providers can limit the number of resources that can be acquired in a period of time. Moreover, due the number of providers, the complexity increases as users have to deal with different Cloud interfaces, pricing schemes and Virtual Machines types.

A Federated Cloud can be used to avoid the unavailability of the resources in case of a Cloud failures and in order to achieve better QoS, reliability and flexibility.

Cloud Federation can be defined as a set of Cloud computing providers, public and private, that voluntarily interconnect their infrastructure through the Internet in order to share resources among each other [13, 14]. In a federated environment, Clouds interact and negotiate the most appropriate resources to execute a particular application/service. This choice can involve the coordination and orchestration of resources that belong to more than one Cloud, which will be used, for instance, in order to execute huge applications. On the other hand, we have Multi-Cloud that denotes the usage of multiple and independent Clouds by a client or service [15].

In this work we use the term Federated Cloud even when the Clouds are not voluntarily interconnected because we are considering the user viewpoint.

In a Federated Cloud context, the usage of an efficient Cloud Broker is essential to abstract the complexity of the Clouds. The role of a Cloud Broker is two fold. First, it provides scheduling mechanisms required to optimize the placement of VM or of the applications among Clouds. Second, it offers a uniform interface with operations to manage the resources independently of a particular Cloud Provider.

The scheduling mechanism in order to optimize the placement of VM or application must take into account the requirements such as the characteristics of the resources, service performance, data locality in order to avoid performance

degradation of the services, the cost, users' QoS constraints and even power consumption. A Cloud scheduler finds an allocation of resources among Clouds Providers that optimizes the user criteria and adheres to some placement constraints.

The Cloud Broker management interface must implement a software layer to translate between generic management operations to specific Cloud API providing a uniform view of the Clouds.

Figure 1 shows the architecture of the proposed Cloud Broker. We assume a hierarchical and hybrid Federated Cloud model. In our model, we have a Provisioning module which is responsible for discovers the resources required to execute an application/service. A Cloud Coordinator to interact with the user and the Clouds. In that case, the Cloud Coordinator works with cooperation of the Provisioning module to control the execution of the applications/services. There is a module to monitor Clouds resources that utilizes QoS parameters or users constraints to take action such as migrating an application/service to another Cloud, consolidating VMs to reduce power consumption, notifying the users through the Provision module about eventually application power leak, or the Cloud Coordinator about the Green Performance Indicators. A Power data logging that is responsible for connect to a power meter and read the power measurements. That module is activated only if there is a power meter connected to the infrastructure. The Power analyzer is responsible for predict the power consumption of the Cloud using the data of the power meter and the data about the Cloud resources usage such as CPU, memory and I/O activity.

The Communication layer is responsible to implement transparent communication among the Clouds. In that layer, generic requests are translated to specific request required by a Cloud provider. This layer is necessary because each Cloud environment may have different limitations, such as maximum request size, distinct request timeouts, distinct commands to interact with the environment and, in some cases, a different execution platform.

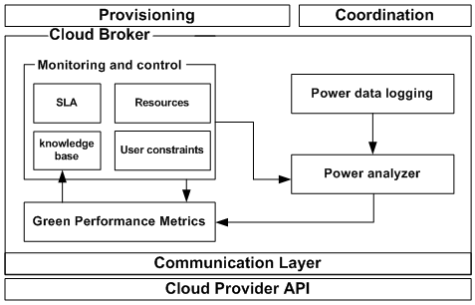

Fig. 1. A Federated Cloud Broker Architecture

4 Conclusion and Future Work

This work presented a hybrid Federated Cloud model that takes into account the power consumption and QoS parameters. For us, a Federated Cloud is an elegant

solution to avoid overprovisioning with a good performance and operational cost. It is also a solution to help data centers to reduce power consumption and its associated carbon footprint.

As future work, we intend to implement the architecture and evaluate it running different applications considering users constraints such as cost, execution time, and Green Performance Indicators (GPIs) related to energy consumption and carbon footprint in a way that the user should pay according to the efficiency of his/her applications in terms of resource utilization and power consumption.

References

1. Greenpeace: Make it green: Cloud computing and its contribution to climate change. Technical report, Greenpeace International (March 2010)
2. Stanford, E.: Environmental trends and opportunities for computer system power delivery. In: 20th ISPSD. (2008) 1–3
3. Hoelzle, U., Barroso, L.A.: The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. M. C. Pub. (2009)
4. Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, G., De Meer, H., Dang, M.Q., Pentikousis, K.: Energy-efficient cloud computing. Comput. J. **53** (2010) 1045–1051
5. Orgerie, A.C., Lefevre, L., Gelas, J.P.: Demystifying energy consumption in grids and clouds. In: GREENCOMP. (2010) 335–342
6. Agosta, G., Bessi, M., Capra, E., Francalanci, C.: Dynamic memoization for energy efficiency in financial applications. In: IGCC. (July 2011) 1–8
7. Chandrakasan, A.P., Brodersen, R.W.: Minimizing power consumption in digital cmos circuits. Proceedings of the IEEE **83**(4) (1995) 498–523
8. Sarood, O., Kale, L.V.: A 'cool' load balancer for parallel applications. In: SC. (2011) 21:1–21:11
9. Mitrani, I.: Service center trade-offs between customer impatience and power consumption. Performance Evaluation **68**(11) (2011) 1222–1231
10. Khazaei, H., Misic, J., Misic, V.B.: Performance analysis of cloud computing centers using m/g/m/m+r queuing systems. TPDS **23**(5) (2012) 936–943
11. Hassan, M.M., Hossain, M., Sarkar, A., Huh, E.N.: Cooperative game-based distributed resource allocation in horizontal dynamic cloud federation platform. Information Systems Frontiers (2012) 1–20
12. Schad, J., Dittrich, J., Quiané-Ruiz, J.A.: Runtime measurements in the cloud: observing, analyzing, and reducing variance. VLDB Endowment **3**(1-2) (2010) 460–471
13. Buyya, R., Ranjan, R., Calheiros, R.N.: Intercloud: utility-oriented federation of cloud computing environments for scaling of application services. In: ICA3PP. (2010) 13–31
14. Celesti, A., Tusa, F., Villari, M., Puliafito, A.: How to enhance cloud architectures to enable cross-federation. In: CLOUD. (2010) 337–345
15. Ferrer, A.J., HernáNdez, F., Tordsson, J., Elmroth, E., Ali-Eldin, A., Zsigri, C., Sirvent, R., Guitart, J., Badia, R.M., Djemame, K., Ziegler, W., Dimitrakos, T., Nair, S.K., Kousiouris, G., Konstanteli, K., Varvarigou, T., Hudzia, B., Kipp, A., Wesner, S., Corrales, M., Forgó, N., Sharif, T., Sheridan, C.: Optimis: A holistic approach to cloud service provisioning. Future Generation Computer System **28**(1) (Jan 2012)