Advanced Energy Management in Cloud Computing

Ahmad Allam, Computer Engineering, Politecnico di MI
Riccardo Chiodaroli, Computer Engineering, Politecnico di MI
Francesco Lunetta, Computer Engineering, Politecnico di TO
Stefano Viganó, Management Engineering, Politecnico di MI
Stefano Ziller, Mathematical Engineering, Politecnico di MI

Principal Academic Tutor:
Antonio Capone, Dept. of Electronics and Information, Politecnico di MI

Other Academic Tutors:
Danilo Ardagna, Dept. of Electronics and Information, Politecnico di MI
Giuliana Carello, Dept. of Electronics and Information, Politecnico di MI
Federico Della Croce, Dept. of Computer Engineering, Politecnico di TO
Marco Mellia, Dept. of Electronics, Politecnico di TO
Michela Meo, Dept. of Electronics, Politecnico di TO

External Tutors:
Massimo Leoni, IBM Italia
Fabrizio Leone, Lutech Spa
Carlo Spinelli, Alcatel Lucent
## Contents

Executive summary 3

1 Introduction 6

2 Green Cloud Project stakeholders 8
   2.1 Cloud Computing 9
      2.1.1 Introduction to the Cloud 9
      2.1.2 Technological overview 10
      2.1.3 Economical overview 14
   2.2 Green Energy and Markets 16
      2.2.1 The Carbon Credit system 16
      2.2.2 Connection with Cloud Computing 17
      2.2.3 The Energy Market 19

3 State of the Art 21
   3.1 Data Center improvements 21
      3.1.1 Virtualization techniques 22
      3.1.2 Control theory approach 23
      3.1.3 Machine learning approach 24
      3.1.4 Utility-based optimization techniques 25
      3.1.5 Cooling infrastructure improvements 26
   3.2 Network improvements 26
      3.2.1 WAN optimization 27
      3.2.2 WAN virtualization 28
      3.2.3 Sleeping mode 28
      3.2.4 Rate switching 30
      3.2.5 Optimization techniques 30
   3.3 Integrating Data Centers and network solutions 31

4 Description of the proposed solution 33
   4.1 Problem description 33
   4.2 Brown model 35
      4.2.1 Parameters 35
      4.2.2 Variables 37
      4.2.3 Objective function 38
Executive summary

Information Technology is developing quickly, with a very significative increase of energy expenses, expected to reach 40% of IT budget by 2012. In this field, Cloud Computing is an emerging paradigm, which is expected to grow exponentially in the next few years.

The Green Cloud project tries to give a solution to the problem of energy consumption in Cloud Computing systems. Indeed, it aims at developing a set of methods for the design of innovative resource allocation policies. Energy and cost savings will be pursued by dynamically allocating computing resources of Data Centers (DC) geographically distributed, trying to find a trade-off with physical constraints and the varying availability and cost of green energy at multiple sites. In this context, our team, whose members come from different engineering backgrounds, tackled the challenge of finding an innovative solution for energy management in Cloud Systems.

Cloud Computing can be described as a system, that users can access via web to obtain computational capacity on demand. The Cloud is composed by a set of Data Centers, that serves users requests, and a network connecting them.

The first step of our work was the definition of a theoretical framework that represents adequately the problem and the current technology: we carried out a broad analysis of what "Green" could mean in a Cloud Computing context. Indeed, since we were considering an emerging paradigm, there were little difficulties in defining what Cloud Computing was. However, none of us had ever thought about energy consumption and pollution applied to such an infrastructure based technology before.

When thinking about making something "greener", the first thought is a technological evolution able to change the way an object or process uses resources. In our case we carried out a systemic study of Cloud Computing (which is summarized in Chapter 1) and we considered a huge amount of different technologies and solutions, including physical DCs, Network components, Virtualization and Routing problems. In any case, we agreed that focusing on a single hardware or software part of the Cloud would not help us finding an innovative solution.

For these reasons, and thanks to our tutors’ support, we focused on a more system-based solution that could integrate Cloud, networking and "Green" concepts. Our problem definition and the results of our systemic study of the solutions proposed in the literature are reported in Chapter 2 and 3, respectively.

The final stage of the our work was therefore to find an original solution to
be developed in the following months of our work. After scouting different approaches, together with the academic tutors we chose to model the Cloud System as a single entity composed both by the DCs and the Network connecting them. Here, the Network serves as a resource for our system, since it allows application requests forwarding, based on specific efficiency policies, aiming at reducing costs or maximizing green energy usage.

With this in mind, we were able to produce an optimization model, which takes care both the Computing part of the Cloud and the Data transfer part. This is also what makes our proposed solution innovative: there is no model in literature able to tackle energy consumption considering both the two main parts of the Cloud jointly.

The solution our group decided to develop is therefore an optimization model aiming to solve the assignment of computational requirements in Data Centers located around the world and the network connecting them, keeping our goal in mind, i.e. the minimization of energy consumption cost and the maximization of green energy. That would benefit both the IT operating firms, that could reduce the energy costs related to their core business, and obviously the environment, that would benefit from the reduction of greenhouse gases emissions. The model and the instances we generated are presented in Chapter 4.

The main decision that our solution takes is to determine if a request direct at the entry Data Center ha to be executed locally or if it has to be redirected to another DC of the network. This problem is solved for a whole day.

Our optimization-based model was developed using the CPLEX software, a high performance automated solver for linear programming problems. A first useful feature of this solution is the reliability of the solving algorithms, allowing to reach optimality and large flexibility in terms of the choice of the scenario (more or less complex) and of model parameters. Overall our model and solution can be adapted and customized to face the needs of the Cloud provider. In fact, we first took into account the most generic scenario, with a large number of request types and different types of servers to elaborate them, then we moved to a more realistic scenario, considering appropriate and realistic parameters and variables, after a wide research.

The main advantage of the solution obtained is the possibility to include in such model all the variables and constraints playing a role in a Cloud system. As a matter of fact, the constraints ensure that the DCs are not overloaded and, at the same time, network capacity is adequate to ensure small delays to the users.

The model was then translated into a real-world scenario, considering a proper geographical dislocation of DCs around the world. A new technical research phase followed, in order to determine the model parameters as close as possible to reality, in terms of servers performance, link bandwidth and so on. Finally, a detailed analysis of the energy cost and the users’ requests in the different region allowed us to set up different instances for our model, determining different results and performance.

As expected, the solution we found is able to redirect the requests in a certain
time bands to the location where the energy cost is lower (mostly night-time locations, due to energy cost fluctuations). As a direct implication, the results show advantages with respect to the approach where DCs and network consumptions are considered separately, which is the usual approach. On the other hand, our model showed great flexibility in scenarios, varying the number of incoming users’ requests and considering different servers performance.

In terms of expense, we experienced savings up to 40% in terms of energy expenses, with respect to the base case, where no request can be forwarded. The main leverages the model exploits are the daily fluctuation of energy prices in the world and the different Data Centers performance. Furthermore, our model, allowing the re-distribution of incoming requests, enables to handle traffic peaks that could not be managed with local computation only.

As a second step, we developed a green model, where we took into account green energy generation, coming from different renewable sources, and varying according to the geographical location and the time band. Thus, the model fosters the usage of renewable energy sources across the system, since we assumed a lower cost with respect to traditional brown energy. Moving the requests were green energy is available allows huge advantages for the environment, in terms of reduction of greenhouse gases emission, which goes far beyond economical savings.

In conclusion, the developed optimization model which can be effectively adopted in practice by any Cloud provider with a private (or eventually, pay-by-use) network which aims at reducing its carbon footprint, as well as breaking energy expenses down.
Chapter 1

Introduction

In this report, we will describe the solution we developed for the Green Cloud project. Before that, however, it may be useful to introduce a few points that have been crucial for the development of the project itself: the problem definition, the requirements and the work method.

The very first day we started working together, we were facing a challenging task, since the problem we were assigned was not a common one. As we know, greenhouse gases emission is becoming a more and more relevant issue in the modern era and the Information Technology and Communication (ITC) industry has always been a relevant part of the solution, since it enables better ways of using energy. Although, we were being asked for something different: we did not have to apply the ITC knowledge to reduce the consumption in an other sector, but rather we had to reduce the consumptions in the ITC sector itself, while maximizing the use of green energy sources. This is a completely new way of thinking about the problem, connected to a widely known issue: ITC industry is growing at a great pace and it will become more and more relevant in any matter, energy consumption included. The important thing to notice is that we were specifically required to work on this kind of matters from a Cloud Computing perspective.

A rising idea in Cloud Computing field is the adoption of a green approach, that is why we are speaking of Green Cloud. By green we mean that it is environment friendly: the aim of this project will be not only to reduce the energy costs for the Cloud provider, but also to reduce the generation of greenhouse gases, by exploiting green energy sources. In this way the Cloud would be more sustainable and environment friendly, especially if the power is generated by using renewable sources of energy.

Moving on with the problem explanation, we were also required to work in strict contact with the academic tutors and other industrial institutions (IBM, Alcatel Lucent and Lutech). The academic tutor immediately started collaborating with us, introducing the problem in details and suggesting some viable approaches. Despite that, we were granted a lot of freedom in the first phase, in order to derive a more creative approach. Industrial partners, on the other hand, helped us validating our approach: we were able to work as we liked and to
present the steps to them in specific meetings that they organized for that purpose. During such events, they were able to give important suggestions and valuable directions to follow, introducing us also some state of the art approaches they were developing.

Therefore, the very first challenge we faced was organizing a work method. We chose an effective approach based on two different phases: a research phase where we split in sub-groups and a second development phase working all together exploiting our multidisciplinary competences. In the first phase, which was mainly research driven, we split into small sub-teams of the same area of competence to greatly increase our knowledge in the main subjects involved. In particular, the sub-teams were three, organized as follows: an economic team, composed by Stefano Viganò, a technology team formed by Francesco Lunetta and Riccardo Chiodaroli and a math team formed by Stefano Ziller and Ahmad Allam. This resulted in a lot of new ideas and in the possibility to write a very technical White Paper [1] overviewing the state of the art in research and in the industrial solutions.

After this, we grouped together with two main objectives: first of all, we had to share the knowledge we had produced with the other members of the teams, granting a unity of vision; after that, we were able to focus on the problem definition and the development of a viable solution. Even though this could seem a simple way of organizing a team-based project, it turned out to be incredibly effective and allowed us to get the best out of the team. Indeed, the amount of multidisciplinary experience that we could bring into the project was high, and the focus on the target of the team in the second phase was the key for achieving our final results.

In the second phase of the project, we worked on the different parts of our solution: from the scenario definition to the mathematical model development, from the parameters setting to the production of the instances. In particular, the development of the theoretical model was leaded by Stefano Ziller, helped by the computer engineers for the technical topics. At the same time, Stefano Viganò proposed the scenario we could use in our model, with a continuous idea exchange among the team members. Once determined the instances we wanted to generate, the whole team provided estimates for all the model parameters, that would allow Francesco Lunetta to run the instances on a Politecnico di Milano virtual machine, thanks to our tutors. Finally, we gathered all the information we had to provide our tutors and external institutions the innovative results of our work, organizing the project deliverables.

Obviously, these two macro phases were organized as mini projects with their deliverables and verification phases. From this point of view, we have to thank the great support of our academic tutors, who followed our mid-term meetings and never stopped to challenge our work.
Chapter 7

Conclusions and future work

As described in the previous chapter, the Green Cloud project solution is built for Cloud providers who wish to reduce their energy expenses as well as the greenhouse gases emissions, in order to optimize the costs linked to the Carbon Credit system. However, the advantages that the use of a Green Cloud solution provides are more than just that. First of all, the management of the entire ICT infrastructure is done by a single firm, with two main benefits: higher economy of scales, that reduce the overall consumption and the possibility to optimize networks and DCs jointly, which is nowadays unexploited. Moreover, Cloud Computing benefits indirectly the firms using it as a service: they can outsource the ICT department, focusing all the efforts on their core business and reducing the energy consumption of operations.

The results we obtained appear very promising. First of all, from an economical point of view, our solution is a way for cutting ICT organizations electrical bills, allowing significant expenses reduction in respect to the base cases, in both the brown and in the green model. As shown in Chapter 6, in fact, our brown models allows savings up to 40% with respect to the current solution used by Cloud providers. This is mainly because our model is able to exploit the time and geographical fluctuations of energy cost, and the most innovative Data Centers, with higher performance servers.

Moreover, from an environmental point of view, the solution we provided allows ICT firms to reduce greenhouse gasses emissions. As a matter of fact, we showed in Section 6.3.1 how our green model is able to exploit all the different green energy sources all over the world exploiting the Cloud network itself, reaching a concrete CO2 reduction.

Finally, our model appears very flexible, not only since it works well with respect to different amounts of daily requests, or different server characteristics, but also because the instances we used can be easily adapted to the needs of a single firm, accordingly to its business and technology.

Further modifications to our model can also be taken into account. First of all, the network topology hypothesis should be relaxed, meaning a detailed description of the underlying network infrastructure should be included within the
model. Up to now, our approach features two parameters related to network topology, the link bandwidth and the number of intermediate routers between the two ends.

Future developments of the model should then account for a non-fully connected topology and different kinds of routing algorithms, as well as introducing transfer delays between links. This is a crucial aspect, since additional model constraints may be introduced, too, with the aim of limiting performance degradation in user experience because of transmission times. Request routing was not included in our approach due to complexity and performance limitations: a model accounting also for routing would not have been solvable within the required time frames. In this sense, it may be interesting to design a secondary optimization problem that, given the first step solution (requested to be moved from one site to another), provides an optimal routing plan for it. Specific routing algorithms, like MPLS, may also promote a similar approach.

In terms of application of performance constraints, instead, Service Level Agreements (SLAs) play an important role for service providers, especially with business of high profile clients. Our model can be extended considering also performance and user experience aspects, which should be definitely added: end users do not accept excessive response time, and would be likely to abort their request if it took more than few seconds to be completed. It could be interesting to model user discontent as a system cost, to be computed in the objective function, or introducing explicit constraints on the end-to-end response time for users’ requests.

Additional test instances could also be defined, by increasing the number of nodes, request and server types, as well as larger traffic shapes. Due to some hardware limitation, in fact, it was not possible to run test case larger than those described in Chapter 6.

However, the results we obtained are encouraging and innovative and we can state that the project has successfully defined an approach for minimizing energy costs in Cloud Computing systems. The approach we followed is a new way of intending Cloud Computing in an environment aware world. Moreover, the number of ICT firms adopting Cloud Systems on their own is increasing at fast pace, showing that our solutions could be applied to an increasing number of real cases. During the New York Interop conference we attended, we could verify that a global approach is becoming more and more interesting, with increasing attention from the whole industry. Cases of firms adopting their own private network are more common now, and it is therefore easier to manage both the network and Data Center aspects of Cloud Computing jointly.

Furthermore, the Green Cloud team is going to present the results obtained during the “Green ICT in Italy: Networks, Cloud and Power Grid” event organized by Politecnico di Milano and Alcatel-Lucent, an industrial partner of the project. The aim of this presentation is to show the feasibility of the approach we followed, in order to obtain visibility on this kind of solutions for further research and improvement. We are looking forward to prove the value of our ASP project!
Bibliography


