Sustainable choices for Cloud Applications: A focus on CO_2 emissions

Cinzia Cappiello, Paco Melia', Barbara Pernici, Pierluigi Plebani, Monica Vitali
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano
Email: [name.surname]@polimi.it

***Abstract*—The widespread adoption of cloud computing is having a big impact on the environment since the energy consumption of data centers and the resulting emissions are significantly increasing. Researchers and practitioners in this field are looking for methods to improve the energy efficiency of data centers and increase the use of green energy sources. In fact, besides the energy consumption, the greenness of a data center can be characterized by the quantity of CO_2 emissions associated to the use of electricity (from a specific energy mix) and/or fuels (e.g. for heating or cooling). In this paper, we propose an approach in which environmental impacts are considered as an important factor for the selection of the cloud site for the deployment of applications. In detail, considering a user perspective and focusing on the assessment of energy consumption and CO_2 emissions, this paper proposes a method to support the users towards greener choices in the deployment of cloud applications.**

I. Introduction

Several goals can act as a motivation for making an organization more sustainable. So far, a relevant role has been held by governmental regulations, which put constraints over the amount of CO_2 emissions that are allowed for an organization. Another relevant driver is the continuous increasing attention of customers towards sustainability, that has brought customers to prefer companies which demonstrated to care about environmental issues. With these motivations, an increasing number of companies is becoming more and more sensitive to the fact that the environmental impact of their systems must be considered. This happens also for IT companies, as data centers and IT systems have a relevant impact over CO_2 emissions. Indeed, a report recently produced by Greenpeace [1] assigns to the IT sector the responsibility for the 2% of the global greenhouse gas (GHG) emissions, pointing out that this percentage is growing. In [2], authors cite data obtained from the EPA (Environmental Protection Agency), stating that GHG have increased from $4.28 \text{x} 10^{13}$ gCO_2 in 2007 to $6.79 \text{x} 10^{13}$ gCO_2 in 2011.

Current trends in research have been often limited on how to reduce the energy demand of data centers. Actually, this is not the only aspect to be considered. As highlighted in the Greenpeace report, an important factor that has to be considered while a company is trying to move towards a more sustainable asset is the *energy production mix* (a.k.a. energy mix): which are, and in which percentage, the sources used to produce the energy consumed by the cloud facilities (e.g., 80% coal, 10% wind, 5% solar, 5% nuclear). In fact, energy sources can not be considered as equal because their environmental impacts (e.g., in terms of CO_2 emissions) are different. In [3], the authors stress this concept by making a distinction between green and non green energy, and by arguing that the impact due to the energy production has to be considered together with the amount of energy consumed.

Focusing on IT solutions, this difference between green and non-green energy must be taken into account, especially in data centers, as it might affect the tasks allocation among the servers in cloud facilities. In fact, the greenness of a data center, defined as the quantity of CO_2 emissions associated to the use of electricity (from a specific energy mix) and/or fuels (e.g. for heating or cooling), can be considered as one of the main driver for the selection of the cloud site for the deployment of applications.

In this respect, the goal of this paper is to propose an approach for a greener deployment of cloud applications, based on the estimation of energy consumption and CO_2 emissions related to these applications. As argued in the rest of the paper, the analysis of the energy mix and the existence of recurring patterns in CO_2 emissions at the national level can be exploited to implement a greener deployment model.

The paper is structured as follows. Section II analyzes previous contributions in order to highlight the novelty of the proposed approach. Such approach is described in Section V and it has been designed considering the reference architecture described in Section III. Section IV defines the assessment methods used to calculate the energy mix in the different cloud sites. Finally, Section VI shows, by using several examples, the effectiveness of the proposed approach. Discussion on future directions of this research is given in Section VII that concludes the work.

II. Related Work

In a cloud environment, several solutions can be adopted to make an IT system more sustainable. Several researcher have been recently trying to use in an efficient way the renewable resources available. It is well known that the main issue of renewable is that they are not constantly available, but their productivity can be dependent on a set of almost unpredictable factors. In [4], authors propose to use a Geographical Load Balancing (GLB) to shift workloads and avoid peak power demands. In order to reach their goals, they propose a prediction algorithm to anticipate future peak demands. The approach is effective principally for a short time prediction, since the error of the algorithm increases with the length of the prediction window. In deciding the

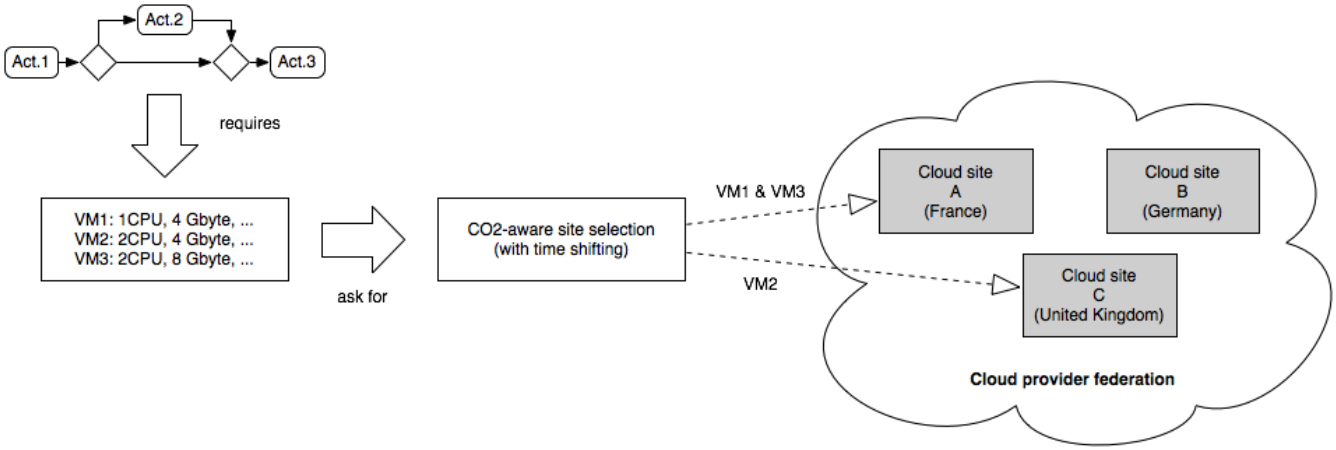

Fig. 1. Reference architecture and scenario

workload allocation, the algorithm takes into consideration renewable energy sources trends, energy storage devices and servers power state. A similar approach is described in [5]. Authors design an algorithm able to reduce the cost due to the energy demand for data centers participating in Coincident Peak Pricing (CPP) programs. These programs charge more when the whole network is in peak of requests. In order to avoid peak load demands, authors use a combination of two techniques: local power generation usage and workload shifting. The algorithm tries to optimize local generation and workload shifting for reducing energy cost. Also in [6], authors highlight the importance of the contribution of green energy sources to promote sustainability in IT systems and propose a load balancing algorithm that takes into account this information when deciding where to allocate the request.

Some scholars face the problem of sustainable business processes focusing their attention on business process reconfiguration. In [7], business processes are redesigned choosing from a set of equivalent fragments for each task. The decision is based on qualitative and quantitative metrics, such as CO_2 emissions, air quality, and damages to fauna and flora. A similar approach is used in [8], where some patterns are defined to design green business processes in a cloud environment. Nine different design patterns are described, which explore different ways to modify the business process toward a greener dimension. As an example, alternative processes can be selected on the basis of their sustainability, some compensation activity can be performed to reduce the environmental effects, or migration can be performed to obtain a greener configuration.

An overview about improving the sustainability of a cloud environment is given in [9], where the software development life cycle in a cloud environment is analyzed from a green perspective. Here, the authors propose a framework in which they identify opportunities for energy efficiency in efficient hardware and software selection, network optimization, scheduling and management of VMs, selection of green energy sources, and efficient data center design.

III. Reference Scenario

The proliferation of cloud platforms has made available to the application developers several infrastructures where to create and deploy their own products. In particular, in this paper, we focus on the IaaS (Infrastructure as a Service) provisioning model, i.e., we assume that the software developers are interested on reserving VMs where the developed application will be executed. Moreover, as a reference scenario we assume that the application that has to be considered is a High Performance Computing (HPC) application. This application can require a set of VMs with different characteristics in terms of number of CPUs, amount of memory, and so on.

With this focus, Fig. 1 gives an overview of the reference architecture and scenario that is considered in this paper. First of all, we are not interested on a single cloud provider but on a federation of cloud providers that can have facilities in different countries. As also discussed in the next section, considering different countries also means considering a different impact on the CO_2 emissions. Indeed, these emissions depend on how much green is the energy production that is used by the application and this, in turn, depends on the energy mix. For the sake of simplicity, to compute the greenness of the energy production in this paper we will refer on the energy mix at national level. For this reason, without loss of generality, we do not consider possible situations in which the cloud facilities have their own autonomous power plants.

Having this cloud infrastructure and this kind of applications, goal of our approach is to find the optimal site where to deploy the requested resources. With respect to the state of the art, this site selection does not only consider the usual constraints on the availability of resources that should be compatible to what it is required. In addition, this selection can also be aware about the environmental impact, measured in terms of CO_2 emissions. Goal of this paper is to focus on this latter aspect analyzing how the energy mix influences this decision in terms of *where* and *when* to deploy the application. In particular, we assume that our approach will be relevant for the manager of the federated cloud, as it is the responsible of identifying the best site where to deploy the application when requested by the final users.

As shown in Fig. 1, the result of the work done by our approach will be the identification of the best way to assign the execution of a task to a given cloud site. For instance, in our example, we obtain that the cloud provider suggests to run the *VM1* and *VM3* in France, while the *VM2* in UK. The three countries included in this example refers to the countries where the three data centers adopted in the ECO2Clouds [1] project are located.

It is worth noting, that even if we are referring in this paper on HPC-like applications, the approach presented hereafter also works for any other kind of applications. We decided to use HPC application as their execution is limited on time and a proposed approach that implies a delay on the execution has more sense. On the contrary, if we considered web-based applications, usually they will be executed for a long period and the possibility to delay the execution of the application results meaningless. Nevertheless, the problem of identify the best site remains relevant also for this class of applications.

IV. EnergyMix Analysis

Greener choices in the deployment of cloud applications should be driven by energy consumption and CO_2 emissions. This means that users should select the cloud site to deploy their applications by considering not only performance but also green requirements. This requires to collect performance metrics, energy metrics and details about the utilization of green sources.

In our approach, the evaluation of the CO_2 emissions is based on the emission factors (gCO2e/kWh) provided by the national grids. Indeed, given the amount of energy consumed, with this factor is possible to compute the amount of CO_2.

Emission factors largely vary from country to country. For example, the three data centers involved in the ECO_2Clouds project are located in France, Germany and United Kingdom, respectively. Some technical reports describe that the country with the lowest carbon intensity is France, whose power generation is mainly based on nuclear plants. Estimated emission factors for France range between 62 [10] and 146 [11] gCO2e/kWh. In contrast, German energy is more carbon-intensive, with emission factor estimates ranging between 629 [12] and 706 [11] gCO2e/kWh. Finally, emission factors for the United Kingdom are estimated to range between 567 [13] and 658 [11] gCO2e/kWh.

As our goal is to deploy an application in a federated cloud environment, the evaluation and estimation of the emission factors is a very important step in our approach. As the factor may vary over time, it is important to know which is the value of the emission factor when the application will run, so that the optimal deployment can occur.

As a first contribution of this paper, we propose different ways to assess the CO_2 emissions.

Looking at the contributions cited in the previous paragraph, we can observe that there are public documents that periodically publish the aggregated emission factors of the different countries in a specific period. In this case, assuming that we know the average power consumption (AP) for a specific site, the energy (kWh) consumed in a specific period can be calculated by multiplying AP by the number of hours in the considered period. CO_2 emissions results multiplying the energy consumed by the emission factor (that is a constant).

[1] http://www.eco2clouds.eu

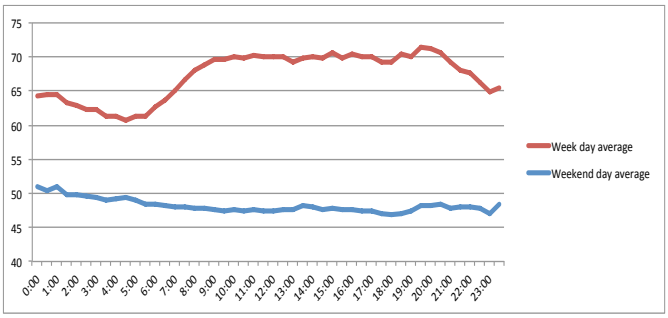

Fig. 3. Trend of the emission factor (gCO2e/kWh) during week days and week ends in France

	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1	0.92	0.67	0.75	0.69
Tuesday	–	1	0.87	0.89	0.83
Wednesday	–	–	1	0.91	0.92
Thursday	–	–	–	1	0.93
Friday	–	–	–	–	1

TABLE I. Correlation indexes of a week measurement

Besides the aggregated emission factors, some countries publish the real time energy mix via public web sites. In this case, the assessment and estimation of CO_2 emissions could be more comprehensive and meaningful. For example, for two of the three countries that we consider in the ECO_2Clouds project real-time energy mixes are available. In particular, France energy mix can be retrieved through the information service *éCO2mix* available on the RTE website [2]. Such service shows electricity demand, electricity generation classified by source and cross-border commercial exchanges (imports/exports). Data are update automatically every 15 minutes. Similar information is available for UK. Real time and historic data about the energy generation in UK are available through the BMRS (Balancing Mechanism Reporting System) website[3]. For this web site data are updated every 5 minutes.

The availability of historical data can be exploited in order to identify regular and/or seasonal pattern that can be used in the deployment of applications. For example, a preliminary analysis of french data of January 2012 revealed the presence of a regular pattern of the emission factors during the week days and another pattern for the weekend days (see Figure 3).

The regularity of the emission factors that characterizes the weekdays has been proven by calculating the correlation indexes among the assessed values of the different week days. As shown in Table I, the correlation indexes are significantly high and it is possible to state that the values gathered in the different days are positively correlated, and thus characterized, by a very similar trend.

[2] http://www.rte-france.com/fr/

[3] http://www.bmreports.com/

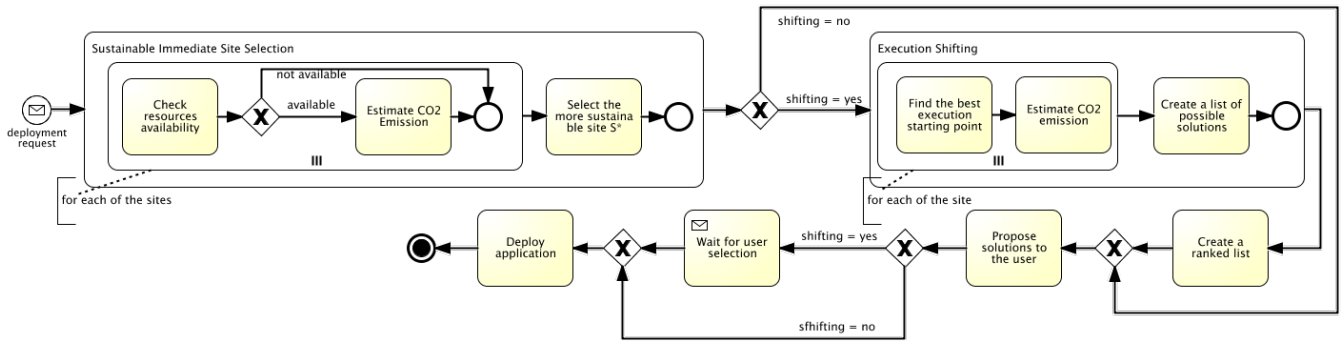

Fig. 2. CO_2-driven site selection

Such trends can be found also in the other months with the difference of seasonal factors. Also the analysis of the British data revealed interesting patterns as better explained in Section VI.

Having these trends, next section will discuss how they can be helpful in driving greener choices for the deployment of cloud applications.

V. CO_2-DRIVEN SITE SELECTION

Assuming that an HPC application, which duration can be estimated, needs to be deployed on one of the available cloud sites. For each of the countries where the cloud providers have their facilities, the energy mix can be estimated according to what it has been discussed in the previous paragraph. Based on that, this section discusses the second main contribution of the paper: i.e., to provide a more green deployment of the application, by making the user aware of the environmental impact of the deployment of its application and by allowing him/her to perform some choices that can reduce the carbon footprint of his(her application. In particular, two complementary approaches are proposed:

- *Immediate site selection*: the approach selects the site according to its carbon footprint, the availability of the resources and the estimated duration of the execution.
- *Execution shifting*: instead of reserving immediately the requested resources, the system proposes different alternative solution that implies a delay of the reservation. This means that the execution of the application will be delayed as well.

Fig. 2, using BPMN notation, illustrates how these two approaches can take place when the user requests for a deployment to the cloud. First of all, the cloud provider enables the users to submit deployment requests for their application. When submitting a request, the user specifies the resources that have to be reserved for the application and an estimated duration of the application (how long the application is going to stay deployed in the cloud infrastructure). The user also specifies his availability in postponing the deployment and the acceptable delay.

Given this information, the cloud provider execute the immediate site selection algorithm and selects the site with the lower estimated CO_2 consumption for an immediate deployment of the application. In case the user has expressed his availability in postponing the execution, the cloud provider performs also the execution shifting algorithm. The output of the algorithm is a list of tuples composed of the name of the site, the delay value, and the estimated CO_2 for the solution. Results obtained are compared with the result of the site selection algorithm and ranked according with the estimated CO_2 emissions. The cloud provider presents to the user the estimation in case of immediate deployment and then the list of the other solution, together with the CO_2 emission reduction and the delay, with an advice about the most convenient combination. The final choice is left to the user.

The two approaches, i.e., immediated and postponed site selection, base their behavior in the patterns observed from the historical value of the CO_2 emission of the considered countries. Details on these approaches follow.

A. Immediate site selection

This approach consists in selecting the site where to deploy an application based on the estimated CO_2 emissions. This estimation is requested as the computation of the carbon footprint has to be done in the near future that covers the execution time of the application. Let's consider a general scenario where a cloud infrastructure is distributed on several sites placed in different countries, each one with its own energy mix and CO_2 emission rate. Some country provide a instantaneous value for CO_2 emission, according to the current production of power. Other provide just a general value that is the average emission value, valid for the whole time. According to this scenario, once a request is received, the infrastructure provider can place the application in the site which is going to be more green than the others for that application.

In Sec. IV we have shown that the emission of CO_2 follows a regular pattern which is dependent from the time of the day, from the day of the week, and from the season of the year. The knowledge of this information for each of the sites that compose the cloud infrastructure is determinant to predict the amount of CO_2 that the request will generate if deployed. The input of the immediate site selection algorithm are the application to be deployed, the resources requested by the application, and an estimated duration in time of the application. The algorithm works as follows:

1) Check the availability of the resources requested by the application on each of the sites S available obtaining a subset of sites S'.
2) For each site in S' predict the CO_2 emission due to the execution of the application given its estimated duration:
   - If instantaneous value are available for the site then estimate the pattern followed by the CO_2 given the past observation.
   - If only a general value is available for the site just multiply the value for the estimated duration of the application.
3) Compare the estimations and select the site S^* with the lowest estimation.

The algorithm for the sustainable immediate site selection can be performed automatically by the cloud provider without involving the user in the decision, since this algorithm just chooses the more sustainable deployment without affecting the performance of the application. On the contrary, the execution shifting algorithm will need an involvement of the user.

B. Execution shifting

In case the execution of the application can be postponed in time, other considerations can be made to further reduce the carbon footprint of the application execution.

In this scenario, we assume that, when the deployment request is sent to the cloud infrastructure, the user can leave to the cloud provider the possibility to postpone the application execution on a more efficient period of time. In this case, the system can investigate better allocation that allows a greater reduction in CO_2 emissions. The user can also specify the maximum allowed delay for the application execution. The execution shifting algorithm base its behavior on the knowledge of the CO_2 patterns discussed in Sec. IV. Given this knowledge, for each site S where the instantaneous energy mix is known the algorithm works as follows:

1) for each site in S find the best execution starting point given the estimated duration of the application and the maximum allowed delay;
2) for each solution, analyze the resulting CO_2 emission values;
3) propose to the user a list of possible deployment solutions with their associated CO_2 emission estimation.

The user is involved in this process since he can decide which is the best solution for his needs by selecting an option from the ranked list.

VI. Validation

In this section we use the knowledge acquired from the analysis of patterns in CO_2 emissions explained in Sec. IV to validate the algorithm described in Sec. V. Through some examples, we describe the potentiality of the proposed approach in reducing emissions. As before, the algorithm is divided into two steps, the sustainable immediate site selection and the execution shifting, that can be analyzed separately.

A. Sustainable immediate site selection validation

This part of the algorithm allows the selection of the best site by comparing different outcomes for an immediate deployment of an application on each of them. As explained in Sec. III, in our example we are considering a cloud environment where three sites are available in three different countries: France, Germany, and United Kingdom. From the analysis of the GHG emissions for each of them, we obtained the plot shown in Fig. 4. Each line in the plot represents the average consumption during a working day and the week end at each of the considered sites. Strong correlations discovered inside a month make us confident that these values can be used as a reference. In the plot the month considered is January 2012, while different results would have been obtained considering other months due to the seasonal variations.

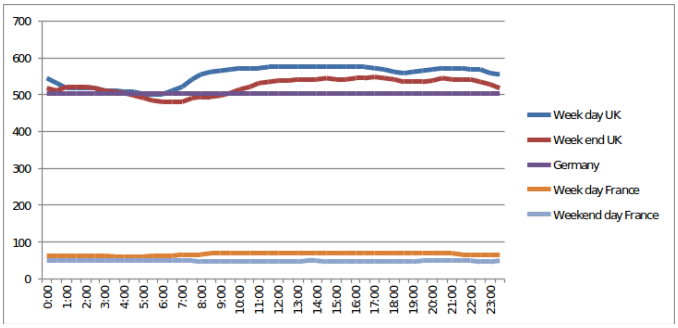

Fig. 4. Trend of the emission factor (gCO2e/kWh) during week days and week end in France, Germany and UK

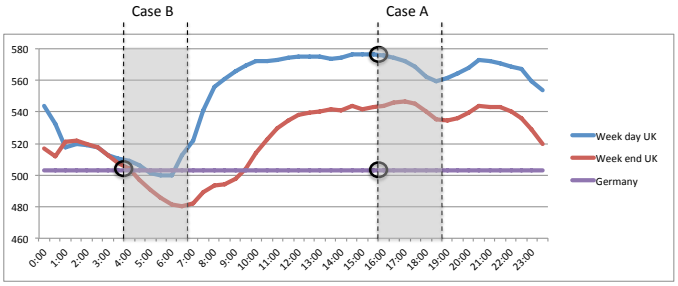

Fig. 5. A detailed view of the trend of emissions for Germany and UK

As can be observed, in our scenario the emissions in Germany are represented using a constant value that is equal to 503 gCO2e/kWh for the specific region where the data center is located. For both UK and France, we rely on the real time values available on the Web and a more detailed analysis can be done. However, in this specific case, France has always a lower emission rate than the other two sites due to its extensive use of nuclear plants. According to this, whenever available, the better choice would always be to deploy the application in the data center located in France. However, resources can also be unavailable for that site, and a comparison is needed for the remaining two sites.

Let us consider a scenario (Case A) where the user asks to deploy an application on Thursday 19th of January at 4:00 p.m.. The application requires to be executed for 3 hours and we estimate an energy consumption of 3 kWh. When the request arrives, resources are available only in UK and

	Delay	Estimated gCO2e	Real gCO2e	Saving (%)
Solution 1	0	209.7	200.3	-
Solution 2	10h	185.4	167.1	16.6%
Solution 3	27h	143.2	140.3	30%

TABLE II. COMPARISON OF EXECUTION SHIFTING OUTCOMES

Germany, so the set of available sites S' is composed of only two sites. A detailed view of the two can be seen in Fig. 5. The estimation for the immediate execution on each site in S' results in 1706 gCO2e for UK and 1509 gCO2e for Germany. According to this, the cloud provider decides to deploy the application in Germany. To demonstrate the effectiveness of the decision we have computed which would have been the actual emission of the application using the available data for the considered date. The real consumption in the UK resulted to be 1677 gCO2e and thus the actual saving is 168 gCO2e. Let us consider the same scenario but when the request arrives at 4:00 a.m. of Saturday 21th. In this case (Case B), estimated emissions are equal to 1469.5 gCO2e for UK and 1509 gCO2e for Germany. The best choice consists in deploying the application in UK where with the real consumption of 1349.5 gCO2e it is possible to save 159.5 gCO2e, even if at the time of the request, Germany had a better emission rate.

B. Execution shifting validation

In this paragraph we validate the second part of the approach, where the customer agrees to postpone the deployment of his application. In order to avoid redundancy we analyze the situation on a single site and for this evaluation we refer to data collected for emissions in France, as shown in Fig. 3. However, the same procedure should be repeated at each site, as discussed in Sec. V-B. Let us consider the same scenario discussed in the previous paragraph where a request arrives on Thursday 19th of January at 4:00 p.m.. The user specifies his availability in postponing the execution with a maximum delay of 48 hours. From an analysis of the trend, the execution shifting algorithm proposes several solutions to the user. The first solution consists in the immediate deployment, with an estimated emission of 209.7 gCO2e. The second solution consists in delaying the execution of 10 hours, by deploying the application on Friday 20th at 2:00 a.m.. In this case, the estimated saving is 24.35 gCO2e. The last solution propose the execution in the week end, starting at 7:00 a.m. of Saturday 21st, with a delay of 27 hours and an estimated saving of 66.5 gCO2e. The user can decide which solution is better according to his needs. In Tab. II the three solutions are compared. The table reports both the estimated and the real values for CO_2 emissions for the three solutions. In the last column it is possible to see the saving in emissions that is obtained when delaying the application deployment. This value is obtained by comparing the effective emissions of the solution to the outcome of the immediate deployment. In this specific example, the algorithm can reduce the emissions of the 30%.

VII. CONCLUDING REMARKS

This paper introduced an approach for considering the CO_2 emissions as a relevant dimension to be considered when applications have to be deployed in a federated cloud. Based on the experience gained in the ECO$_2$Clouds project, the paper has presented an approach for analyzing the energy mix to discover patterns that can be exploited in the deployment phase. Moreover, the paper also introduced a site selection algorithm that considers the CO_2 emissions in two cases: an immediate deployment and a delayed deployment. Validation scenario based on real data publicly available on the energy mix in France and UK shows how energy savings can be obtained following a particular deployment strategy.

Next steps in this research will take into account the impact of an energy source from cradle to cradle. This means that the nuclear sources will not be considered as a green source. Indeed, even if it has limited CO_2 emissions, considering also the building and the decommissioning this value will increase. Moreover, also the risk associated to a given energy source will be considered. Yet, considering the nuclear power, users might tend to avoid to choose cloud sites fed by this kind of sources as they perceive the associated risk of contamination.

ACKNOWLEDGMENT

This work has been supported by the ECO$_2$Clouds project (http://eco2clouds.eu/) and has been partly funded by the European Commission's IST activity of the 7th Framework Programme under contract number 318048.

REFERENCES

[1] G. Cook, "How Clean is Your Cloud?," tech. rep., Greenpeace International, April 2012.

[2] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in Computers*, vol. 82, no. 2, pp. 47–111, 2011.

[3] J.-M. Pierson, "Green Task Allocation: Taking into Account the Ecological Impact of Task Allocation in Clusters and Clouds," *Journal of Green Engineering*, vol. 1, no. 02, p. 11761, 2011.

[4] Z. Abbasi, M. Pore, and S. Gupta, "Impact of workload and renewable prediction on the value of geographical workload management," in *Second International Workshop on Energy Efficient Data Centers (E2DC)*, 2013.

[5] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation," *Performance Evaluation*, vol. 70, no. 10, pp. 770–791, 2013.

[6] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," *In Proc. of Sigmetrics 2011*, 2011.

[7] K. Hoesch-Klohe and A. Ghose, "Carbon-Aware Business Process Design in Abnoba," in *ICSOC*, pp. 551–556, Springer, 2010.

[8] A. Nowak, T. Binz, C. Fehling, O. Kopp, F. Leymann, and S. Wagner, "Pattern-driven green adaptation of process-based applications and their runtime infrastructure," *Computing*, vol. 94(6), pp. 463–487, 2012.

[9] N. S. Chauhan and A. Saxena, "A Green Software Development Life Cycle for Cloud Computing," *IT Professional*, vol. 15, no. 1, pp. 28–34, 2013.

[10] ADEME, "Guide des facteurs dmissions, Version 6.1. Chapitre 2, Facteurs associs la consommation directe dnergie," tech. rep., 2010.

[11] European Commission, "European Commission (2010) How to develop a Sustainable Energy Action Plan (SEAP)," tech. rep., 2010.

[12] Umweltbundesamt, "Entwicklung der spezifischen Kohlendioxid-Emissionen des deutschen Strommix 1990-2010 und erste Schtzungen 2011," tech. rep., 2012.

[13] Department for Environment, Food and Rural Affairs, "Guidelines to Defra/DECC's GHG Conversion Factors for Company Reporting: Methodology Paper for Emission Factors," tech. rep., 2012.