Classification of Techniques for Energy Efficient Load Distribution Algorithms in Clouds – A Systematic Literature Review

Matthias Splieth, Frederik Kramer, Klaus Turowski[1]

Abstract

Cloud Computing has become an important driver for IT service provisioning in recent years. In addition to its associated benefits for both customers and IT service providers, cloud computing also comes along with new challenges. One of the major challenges for providers is to reduce the energy consumption, since today already more than fifty percent of operational costs in data centers account for energy. A possible way to reduce these costs is to distribute load in terms of virtual machines within the data center. Developing algorithms for this purpose has been a topic of recent research. In order to capture the state of the art of energy efficient load distribution in clouds, this paper presents a structured literature review on load distribution algorithms that aim to reduce the energy consumption in data centers for cloud computing. The algorithms are reviewed in terms of their type, the evaluation method and its potential side effects in terms of drawbacks.

1 Introduction

In recent years, IT has become an important factor for the increasing energy demand. Especially with the rise of cloud computing, large data centers are being built. These consume huge amounts of energy which cause high operational costs. Current studies have shown that the energy consumption of data centers increased by 56% from 2005 to 2010 [1]. Given that more than fifty percent of the overall costs in data centers account for energy, it is eligible for data center operators even to slightly decrease the energy consumption since this can have a significant impact on their profitability [2]. While data centers have several energy consuming components, a major source of wasting energy is the inefficient usage of computing resources [3]. An important method that can improve the utilization of resources and at the same time reduce the overall energy consumption is to load distribution within a data center by migrating virtual machines [2]. Since in clouds the resources used by customers are scaled according to their current demand, distributing load is very important. Thereby, load distribution is one of the major challenges that cloud data centers need to face [4]. In recent years, numerous algorithms for distributing load in clouds have been developed, many of these with the aim to reduce energy consumption. These algorithms apply different techniques in order to save energy: as an example, some algorithms use techniques such as rule-based migration of virtual machines [5] while other algorithms rely on genetic programming [6]. Hence, there are multiple approaches to reduce energy consumption in a data center. An analysis of such algorithms can help to make their results comparable and to identify potential side effects, for example in terms of negative effects on the availability of a service. In addition, the need for the design and the evaluation of algorithms for energy efficient load distribution in clouds can be identified in order to point out gaps in current research.

1 Otto von Guericke University Magdeburg, Faculty of Computer Science, Very Large Business Applications Lab, Universitätsplatz 2, 39016 Magdeburg, Germany (matthias.splieth,frederik.kramer,klaus.turowski)@ovgu.de

1. Research Question

In order to be able to classify algorithms that can be applied to reduce energy consumption in data centers, the first research question that is to be answered in this paper is:

> **RQ1**: Which types of load distribution algorithms are used in order to reduce the energy consumption in data centers for cloud computing?

Since there is no standardized methodology for the evaluation of algorithms in terms of improvements in energy consumption, the question arises how potential improvements are estimated. Thus, the second research question investigated in this paper is:

> **RQ2**: How are the improvements in energy consumption evaluated?

Load distribution can have drawbacks. For example, the migration of a virtual machine between two different physical servers can affect the performance and hence result in higher response times. This in turn can lead to violations of service level agreements (SLA) and result in contractual penalties. Investigating potential side effects is thus an important factor when analyzing algorithms. Therefore, the third research question aims to identify such side effects:

> **RQ3**: Have side effects been investigated and, if yes, which side effects have been described?

In order to answer those research questions, a structured literature review is conducted.

2 Related Work

There is only little regarding the systematic analysis of energy efficient load distribution algorithms in clouds. In [7], a survey and analysis on resource scheduling algorithms is presented. However, energy efficiency is only marginally considered since it is not the focus of the paper. In publications that deal with the development of new algorithms, comparisons are made with other algorithms in general. However, these usually do not include a systematic analysis of other algorithms [8]. Some paper do provide more detailed information about other algorithms, but focus on very specific types of algorithms, such as for Ant Colony Optimization in [9].

3 Literature Review

According to [10], a literature review consists of the four steps *material collection*, *descriptive analysis*, *category selection* and *material evaluation*. The material collection defines the material that is to be collected. The descriptive analysis provides an evaluation of formal aspects of the material. The category selection defines structural dimensions that are used to analyze the collected material. The material evaluation comprises the evaluation of the material and an interpretation of the results. Before the review is carried out, the delimitations of this study are presented.

2. Delimitations

This literature review is a representative study that aims to identify relevant research outcomes [11] with respect to the research questions defined in section 1. Therefore, a set of criteria was defined to be able to select relevant publications. In order to be recognized as relevant, a publication needs to meet all inclusion criteria (I) while not meeting any of the exclusion criteria (E). Table 1 lists all criteria that were used to determine the relevancy of a publication.

Inclusion Criteria		Exclusion Criteria	
I 1	The publication is written in English.	*E 1*	The algorithm requires a specific hardware configuration.
I 2	The paper is published in a proceeding or journal.	*E 2*	The algorithm is designed for a specific problem, for example MapReduce.

I 3	The publication describes an algorithm that aims to reduce energy consumption.	*E 3*	The publication does not have an emphasis on energy efficiency.
I 4	The algorithms' setting is a cloud computing environment.	*E 4*	The publication does not clearly describe the functionality of the algorithm.
I 5	The publication is in the form of a completed research paper.	*E 5*	The publication does not clearly describe the evaluation of the results.
		E 6	The algorithm requires all information a priori in order to distribute load.

Table 1: Inclusion/Exclusion Criteria

3. Material Collection

In order to identify relevant research outcomes, a structured keyword search was conducted. The applied search term, the time frame investigated in the context of the search and the queried databases are listed in Table 2Table 1.

Parameter	Value
Search term	energy efficiency *OR* energy efficient *OR* energy conservation *OR* energy aware *OR* green it *OR* green computing *OR* green ict *OR* energy saving *OR* sustainable *OR* sustainability *OR* energy consumption *OR* power consumption *OR* power management *OR* energy costs *AND* (load balancing *OR* load distribution) *OR* ((vm *OR* virtual machine) AND (migration *OR* allocation *OR* provisioning *OR* scheduling *OR* placement)) *OR* live migration *OR* capacity management *OR* resource management *AND* cloud computing *OR* infrastructure as a service *OR* software as a service *OR* platform as a service *OR* iaas *OR* sass *OR* paas *OR* cloud *OR* data center *OR* data centre
Time frame	2008 – 2014
Databases	ACM Digital Library, AIS Electronic Library, Computing Research Repository, Directory Of Open Access Journals, Ebsco Host, ScienceDirect, Emerald, IEEE Xplore Digital Library, JSTOR, Oxford Computer Journal Database, Palgrave Macmillan, SpringerLink, Taylor & Francis Online, Wiley Online Library, WISO

Table 2: Parameters of the literature review

The period of search covers the years from 2008 to 2014. The beginning of the period was set to 2008 since this is the year in which *Amazon* introduced its cloud platform *EC2* and thus made cloud computing a popular topic. The major databases (according to [12]) have been selected for the search. The applied search term as shown in Table 2 is split into three parts. Each part describes a term that is relevant for this research including various synonyms. The first part consists of terms regarding energy consumption and energy savings. The second part includes synonyms for load distribution in terms of migrating or placing virtual machines. The last part limits the search to cloud environments. Within the individual parts, the terms are connected with a Boolean *OR*. The three parts however are connected with a Boolean *AND*. Terms within the search query that consist of two or more words (such as "energy efficiency") are treated as phrases. The querying with the search term was restricted to abstracts. If a database did not support this type of search, a full text

search was applied. In order to decide whether a publication is relevant, its abstract was read. In the case that only a full-text search was possible (such as for SpringerLink), a title filtering was conducted before reading the abstract. If the relevance of an article cloud not be determined clearly, the entire article was read.

The search query returned 1910 results. Their abstracts were tested according to the inclusion and exclusion criteria that are listed in Table 1. After this first refinement, 125 publications remained. Subsequently, these articles were completely read and checked according to the criteria. As a result, 80 publications remained that are relevant for this literature review.

4. Descriptive Analysis

The set of relevant publications comprises 80 papers. The distribution of these papers over the search period is presented in Figure 1. While there are only few publications located in the beginning of the search period, there is a strong increase in the number of publications in the subsequent years. This increased research attention indicates the importance of energy efficient load distribution in data centers. In addition, the high number of publications suggests that this topic is attractive to the scientific community and still has open questions.

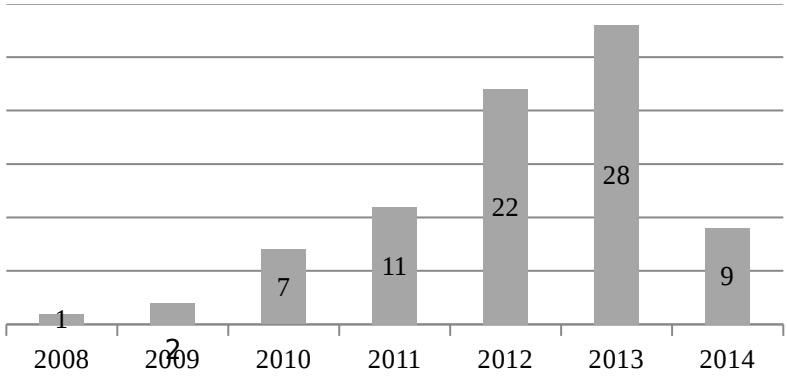

Figure 1: Number of publications per year over the reviewed time period

The majority of the papers was published in proceedings (54 publications), the smaller part in journals (26 publications). These 80 publications spread over 63 different outlets. Therefore, leading publication outlets are hard to locate. In fact, this relatively equally distribution shows the wide acceptance of the topic in research.

5. Category Selection and Material Evaluation

Since this paper aims to examine algorithms for energy efficient load distribution in clouds as well as the applied evaluation method and side effects of the algorithms, the categories for the material evaluation are derived inductively while reviewing the material [10]. These categories are made up by the type of algorithm, the evaluation method and the investigated side effects. A total of 28 different types of algorithms was identified and then grouped into 11 categories as shown in Figure 2. The grouping was conducted based on the field of the algorithms. For example, the topics Reinforcement Learning [13] and Fuzzy Q Learning [14] appeared in the collected material and both belong to the field of machine learning. Therefore, the group *Machine Learning* was build.

Static algorithms include simple threshold-based [15] and rule-based approaches [5]. All integer programming approaches such as Boolean Integer Programming [16] or Mixed Integer Programming [17], are grouped in **-Integer Programming*. *Machine Learning* refers to algorithms that use Reinforcement Learning, Fuzzy Q Learning, or Learning Automata. **-Fit* comprises of algorithms that use Best Fit [18] or First Fit [19] approaches or modifications of these. In *Bio-*

inspired Computing, algorithms are aggregated that try to imitate their biological counterparts, for example genetic algorithms [20] or algorithms that apply ant colony optimization [21]. The category *Others* represents independent types of algorithms that only appeared once in the collected material. An example from this category is an algorithm that uses Lyapunov optimization [22] or an algorithm that uses Space Partitioning in order to distribute load [8]. The frequency of algorithms exceeds the number of publications because some papers, for example, introduce more than one algorithm [23] or combine different types of algorithms [24].

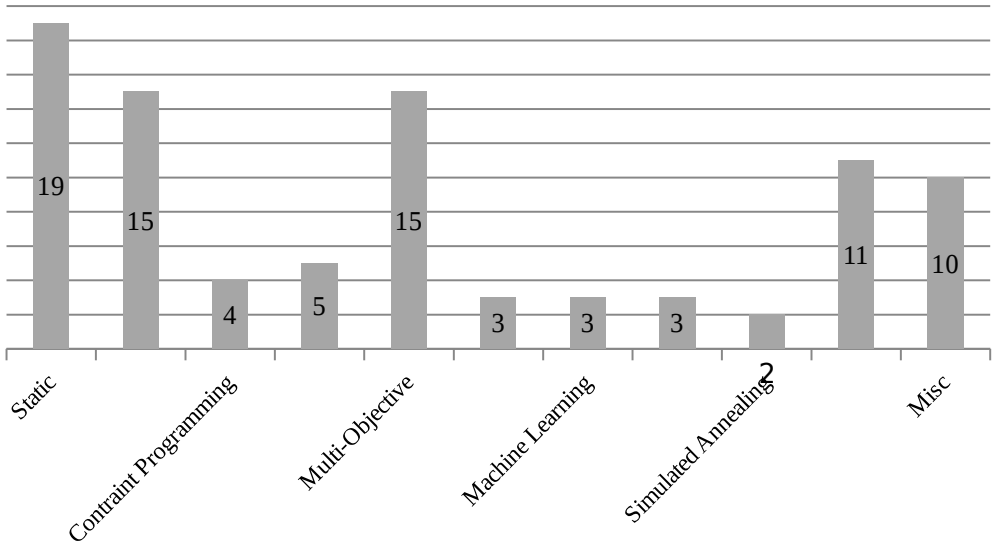

Figure 2: Frequency of Algorithms in publications grouped by the type of algorithm

The majority of the publications seek to provide better solutions for the bin-packing-problem, meaning to provide a solution that is more energy efficient for placing virtual machines in a data center. While algorithms which were published at the beginning of the investigated time frame rely on simple techniques such as rule-based approaches [5] in order to do this, the algorithms are getting more complex over time, such as in [25] or [22]. In order to solve the optimization problem, the publications use or adapt ideas and approaches from other domains, such as mathematics or economics. Besides static algorithms, Linear Programming approaches and Multi-Objective optimizations are the most common implementations. In recent years, Bio-inspired computing is gaining more attention with respectively three publications in each 2012, 2013 and 2014. Most of the investigated algorithms are working reactively. This means that they only react to specific situations, such as an overloaded server, in order to start a load distribution procedure [26]. It would be better however if the algorithms prevented such situations proactively, for example by using prediction methods in order to distribute the load according to a prediction, such as presented in [27]. This could help to use overbooking while simultaneously not decreasing the service level. However, only few algorithms use prediction approaches (confer figure 2).

With regard to the second research question, namely how the improvement of the energy consumption is measured, the analysis showed that the majority of the algorithms are evaluated by simulation (67 publications). A much smaller proportion used test environments for evaluation (14 publications). Only four publications conducted a purely mathematical evaluation of their proposed algorithm(s). Only one paper did not carry out an evaluation. The sum of evaluated publications exceeds the total number of publications since some algorithms were evaluated using multiple evaluation methods, such as in [28]. A major difference in the evaluation can be found in terms of the quality of the simulation. The majority of the authors do not use real data in order to strengthen the simulation results. Additionally, most algorithms are tested only with respect to one specific

workload. Since the workloads can be unpredictable [29], it does not make sense to test against one specific workload only. Testing against a single workload shows the potential effect of an algorithm in a specific case, but does not allow making conclusions about the general quality of the algorithm. Precisely, only 26% of the papers, such as [29] or [25], explicitly consider different types of workflows and thus demonstrate the suitability of their algorithm. Furthermore, only about 50% among these 26% use real workload data in order to evaluate their developed algorithm.

With regard to the third research question, the literature review has revealed that by far not all papers do consider side effects. In total, 33 publications at least mention side effects and claim that their investigation is highly relevant − in general, they refer to terms such as “SLA-violations” in order to describe side effects. In fact, only 28 papers seriously investigate side effects. The most investigated aspect in terms of (negative) effects of algorithms is performance (investigated in 14 papers). But also response time (in 6 papers), availability (in 6 papers) and throughput (in 2 papers) are investigated. However, the investigations of these aspects are hardly comparable with each other because they mostly apply different, often self-defined metrics in order to determine the effects. In some cases, it is even questionable if the applied metrics are related to the side effect at all. For example, Khosravi et al. define SLA-violations as the number of rejected virtual machine requests [30]. This is not a suitable metric for SLAs or even for the availability of a service, which is apparently meant by this definition. Thus, this important aspect seems to be still ambiguous.

4 Discussion

The conducted literature review revealed a few weak points in in the research on load distribution in clouds. First of all, many papers do only consider low-load situations. For cloud environments, this is a problematic assumption since in clouds, the servers are usually higher utilized or the workloads are even unpredictable [29]. Of course a low-load situation offers more potential for energy savings, but may not be applicable in real-world scenarios.

Secondly, it can be stated that the simulation-based evaluations of the individual papers are hardly comparable. A total of 69 papers used simulation in order to evaluate their contributions. Out of these, 37 used a proprietary development. Since it is often not clear how these proprietary tools are constructed or even how they estimate energy consumption, their results are difficult to compare. Seven papers did not use a proprietary tool, but instead each used a tool that just occurred once in the collected material. Therefore, their results are also difficult to compare among each other. In total, 25 papers used *CloudSim* [31] for the evaluation. While these results are comparable with each other, the problem hereby is that *CloudSim* has not been developed for ascertaining the energy consumption in clouds. There are indeed components included for this purpose, but the implementation is very rudimentary: servers are the only energy-consuming components of a data center and in the servers, only the CPU consumes energy. Having a look at the energy-consuming components of a data center (such as presented by Jing et al. in [32]), this approximation is too simplistic. The validity of the results evaluated with *CloudSim* in terms of the overall energy consumption of a data center is thus at least questionable. Thirdly, the paper does not go due to the simulators used on other effects, such as, for example, additional energy savings due to a better heat generation. In particular, these thermal aspects are generally not considered.

Another problematic aspect is that in many evaluations, a homogenous infrastructure for the whole data center is assumed. In practice, it is unlikely that the entire infrastructure of the data center is completely homogeneous. This may be a realistic assumption for single clusters within a data center, but not for an entire data center. Therefore, this assumption cannot be made in practice.

Summing up, it can be stated that these aspects that were not sufficiently considered should be taken into account in future research.

5 Conclusion

In order to capture the current state of the art of energy efficient load distribution algorithms in cloud data centers, a structured literature review was conducted. The number of results indicates that this is a highly relevant topic in current research. Thereby, the interest in energy efficient load distribution has significantly increased in recent years and it is likely to expect that this trend will continue. The advantage of the investigated algorithms is their independency of technical aspects of data centers. The algorithms can work with different hardware configurations while not needing any information about the cooling status and nevertheless still reaching their goals. In fact, the results of the identified research outcomes indicate that the overall energy consumption can be reduced in most cases. Unlike other methods that can be used to save energy, such as Dynamic Voltage and Frequency Scaling or partial shutdowns, load distribution algorithms implicate an aspect of generality. Since data centers are very diverse, this generality makes the idea appealing.

However, the literature review has also shown that several aspects have not been sufficiently considered in research yet. Although the review showed that the algorithms become more complex over time and augur higher energy savings, the analysis also revealed that the evaluation in terms of simulation needs to be improved since there are still too many aspects left out. For example, the simulation tool for evaluation that is most commonly applied, *CloudSim*, only models the CPU as energy consuming component in a data center. Evaluation tools that do not rely on *CloudSim* are often confronted with the problem that their results are not replicable. Therefore, both the comparability of the different results and the transferability of the results in practice are doubtful. It is time to combine the research to build a fully-functional prototype, implementing an architecture that covers all aspects and not just elements. This must be addressed in future research.

References

[1] J. Koomey, “Growth in Data center electricity use 2005 to 2010,” Analytics Press, Oakland, Aug. 2011.

[2] A.-C. Orgerie, M. D. De Assuncao, and L. Lefevre, “A Survey on Techniques for Improving the Energy Efficiency of Large Scale Distributed Systems,” *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–35, Dec. 2014.

[3] L. A. Barroso and U. Hölzle, “The Case for Energy-Proportional Computing,” *IEEE Computer*, vol. 40, pp. 33–37, Dec. 2007.

[4] Q. Zhang, L. Cheng, and R. Boutaba, “Cloud computing: state-of-the-art and research challenges,” *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, Apr. 2010.

[5] C. Kleineweber, A. Keller, O. Niehörster, and A. Brinkmann, “Rule-Based Mapping of Virtual Machines in Clouds.,” in *Proceedings of the 2011 IEEE 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP ’11)*, 2011, pp. 527–534.

[6] B. Kantarci, L. Foschini, A. Corradi, and H. T. Mouftah, “Design of energy-efficient cloud systems via network and resource virtualization,” *International Journal of Network Management*, 2013.

[7] C. T. Lin, “Comparative Based Analysis of Scheduling Algorithms for Resource Management in Cloud Computing Environment,” *International Journal of Computer Science and Engineering*, vol. 1, no. 1, pp. 17–23, 2013.

[8] W. Huang, X. Li, and Z. Qian, “An Energy Efficient Virtual Machine Placement Algorithm with Balanced Resource Utilization,” in *Proceedings of the 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS ’13)*, 2013, pp. 313–319.

[9] A. E. Keshk, “Cloud Computing Online Scheduling,” *IOSR Journal of Engineering*, vol. 4, no. 3, pp. 7–17, Mar. 2014.

[10] S. Seuring and M. Müller, “From a literature review to a conceptual framework for sustainable supply chain management,” *Journal of Cleaner Production*, vol. 16,

no. 15, pp. 1699–1710, Oct. 2008.

[11] J. vom Brocke, A. Simons, B. Niehaves, K. Reimer, A. Cleven, and R. Plattfaut, “Reconstructing the giant: On the importance of rigour in documenting the literature search process.,” in *European Conference on Information Systems*, 2009, pp. 2206–2217.

[12] J. Hintsch, “ERP for the IT Service Industry: A Structured Literature Review,” in *Proceedings of the Nineteenth Americas Conference on Information Systems 2013 (AMCIS ’13)*, 2013.

[13] W. Yan, C. Lin, and S. Pang, “The Optimized Reinforcement Learning Approach to Run-Time Scheduling in Data Center.,” in *Proceedings of the 2010 IEEE 9th International Conference on Grid and Cooperative Computing (GCC ’10)*, 2010, pp. 46–51.

[14] S. S. Masoumzadeh and H. Hlavacs, “Integrating VM selection criteria in distributed dynamic VM consolidation using Fuzzy Q-Learning,” in *Proceedings of the 2013 IEEE 9th International Conference on Network and Service Management (CNSM ’13)*, 2013, pp. 332–338.

[15] C.-C. Lin, P. Liu, and J.-J. Wu, “Energy-efficient Virtual Machine Provision Algorithms for Cloud Systems,” in *Proceedings of the 2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC ’11)*, 2011, pp. 81–88.

[16] B. Yin and L. Lin, “Energy reducing dynamic multi-dimensional resource allocation in cloud data center,” in *Proceedings of the 2012 IEEE 14th Asia-Pacific Network Operations and Management Symposium (APNOMS ’14)*, 2012, pp. 1–4.

[17] Z. Abbasi, T. Mukherjee, G. Varsamopoulos, and S. K. S. Gupta, “DAHM: A Green and Dynamic Web Application Hosting Manager Across Geographically Distributed Data Centers,” *Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 8, no. 4, pp. 34:1–34:22, Nov. 2012.

[18] D. M. Quan, A. Somov, and C. Dupont, “Energy Usage and Carbon Emission Optimization Mechanism for Federated Data Centers,” in *Proceedings of the First International Conference on Energy Efficient Data Centers (E2DC’12)*, 2012, pp. 129–140.

[19] I. Rodero, H. Viswanathan, E. K. Lee, M. Gamell, D. Pompili, and M. Parashar, “Energy-Efficient Thermal-Aware Autonomic Management of Virtualized HPC Cloud Infrastructure,” *Journal of Grid Computing*, vol. 10, no. 3, pp. 447–473, Sep. 2012.

[20] F. F. Moghaddam, M. Cheriet, and K. K. Nguyen, “Low Carbon Virtual Private Clouds,” in *Proceedings of the 2011 IEEE 4th International Conference on Cloud Computing*, 2011, pp. 259–266.

[21] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, “A Multi-objective Ant Colony System Algorithm for Virtual Machine Placement in Cloud Computing,” *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, Dec. 2013.

[22] S. Ren and Y. He, “COCA: Online Distributed Resource Management for Cost Minimization and Carbon Neutrality in Data Centers,” in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC ’13)*, 2013, pp. 1–12.

[23] D. Borgetto, M. Maurer, G. Da-Costa, J.-M. Pierson, and I. Brandic, “Energy-efficient and SLA-aware Management of IaaS Clouds,” in *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy ’12)*, 2012, pp. 1–10.

[24] J. Dong, H. Wang, X. Jin, Y. Li, P. Zhang, and S. Cheng, “Virtual Machine Placement for Improving Energy Efficiency and Network Performance in IaaS Cloud,” in *Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops*, 2013, pp. 238–243.

[25] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch, “AutoScale: Dynamic, Robust Capacity Management for Multi-Tier Data Centers,” *ACM Transactions on Computer Systems*, vol. 30, no. 4, pp. 1–26, Nov. 2012.

[26] H. Cambazard, D. Mehta, B. O’Sullivan, and H. Simonis, “Constraint Programming Based Large Neighbourhood Search for Energy Minimisation in Data

Centres.," in *Proceedings of the 10th International Conference on Economics of Grids, Clouds, Systems, and Services*, 2013, pp. 44–59.

[27] X. Li and M. Zheng, "An Energy-Saving Load Balancing Method in Cloud Data Centers," in *Proceedings of the 2013 International Symposium on Information Technology in Medicine and Education (ITME 2013)*, 2014, pp. 365–373.

[28] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch, "AutoScale: Dynamic, Robust Capacity Management for Multi-Tier Data Centers," *Future Generation Computer Systems*, vol. 36, no. 0, pp. 237–256, 2014.

[29] T. C. Ferreto, M. A. S. Netto, R. N. Calheiros, and C. A. F. De Rose, "Server Consolidation with Migration Control for Virtualized Data Centers," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1027–1034, Oct. 2011.

[30] A. Khosravi, S. K. Garg, and R. Buyya, "Energy and Carbon-efficient Placement of Virtual Machines in Distributed Cloud Data Centers," in *Proceedings of the 19th International Conference on Parallel Processing (Euro-Par'13)*, 2013, pp. 317–328.

[31] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," in *Proceedings of the International Conference on High Performance Computing & Simulation*, 2009, pp. 1–11.

[32] S.-Y. Jing, S. Ali, K. She, and Y. Zhong, "State-of-the-art research study for green cloud computing," *The Journal of Supercomputing*, vol. 65, no. 1, pp. 445–468, 2013.