Load distribution strategies for a sustainable IT resources management

Abdulrahman Nahhas[1], Sascha Bosse[1], Klaus Turowski[1]

[1] Very Large Business Applications Lab, Faculty of Computer Science, Otto-von-Guericke-University, Magdeburg, Germany
{firstname.lastname}@ovgu.de

Abstract. The rapid transformation and virtualization strategies of IT resources are evolving all possible fields of IT-industries. With the introduction of cloud computing model, the structural complexity of IT-infrastructures is radically increasing. The engineering process of a system landscape itself is not anymore the central task to optimize but also to efficiently utilize that enormous system landscape. Energy-costs are constantly increasing and to keep low-cost service, IT service providers are willing to investigate their energy efficiency. Moreover, they are recently obliged to reduce their CO_2 emissions under global and governmental pressures. In this paper, we present a holistic view of the virtual machines placement problems, their challenges and currently used solution methodologies, which target reducing energy consumption and aim for a sustainable IT resources management. In addition, we present a case study to demonstrate the role of load distribution strategies in sustainably managing IT resources to reduce energy consumption.

Keywords: Load distribution strategies, sustainable IT resources management, Virtual Machines Placement problem, IT energy consumption.

1 Introduction

The rapid transformation and virtualization strategies of IT resources have been radically evolving all possible fields of IT markets and industries. Nowadays, everything is or might be shifted to the cloud and proposed in the market for different customer sectors as services based on the model of cloud computing. Cloud computing, by definition, is a model for providing a rapid on-demand network access to a shared pool of different computing resources that can be with the minimum administrational effort provisioned and supplied for customers with different needs [1]. Infrastructure, platform and applications are now delivered as services based on the concept of cloud computing. This model triggered a new era of innovation in the IT- industry and raised the competition to deliver better services with the minimal costs.

Using the concept of the cloud allows IT service providers to employ their resources more in new innovations rather being obliged to install and set up the necessary basic hardware and software infrastructure for clients [2]. However, this model has also introduced new challenges in addition to the normal system landscape engineering ones and spotted many new obstacles in dealing with that rapid growth of IT system

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

landscapes. The engineering process of the system landscape itself is not anymore the central task to optimize but also to efficiently utilize that system landscape. In other words, reducing the tremendous costs and investments in the IT infrastructure by the IT service providers is not anymore the only concern but rather reducing the energy power consumption and the associated operational costs of that infrastructure. However, the elasticity characteristic of the cloud has also radically raised the complexity of those systems and predicting the current required computational power is rather impossible due to the high heterogeneity of the workload patterns [3].

A recent study revealed, that the electricity consumption of data centers has reported worldwide a tremendous growth in the last years and for instance, between 2005 and 2010 the electrical power consumption by data centers has doubled [4]. Other studies have stressed on the electricity consumption and its large proportion of the overall operational costs of IT services providers, which is estimated to exceed 50 % of the overall operational costs [4, 5]. The central component in these calculations is the active servers, which consume the most electricity in data centers in comparison to other cooling and ventilation components [4]. Some studies showed that a critical number of functional servers in data centers in the USA, for instance, has never been utilized and rough calculations suggest that a total waste of over US- $ 19 billion is encountered per year with a total of 11 million tons of CO_2 emissions [6]. Holding the Service Level Agreement (SLA) has been the most common target objective for many IT service providers. Seizing opportunities to reduce costs and enhancing the enterprise global image under sustainable IT resources management are recently the keys to obtaining a better position in the market in addition to the quality of provided services. Energy costs are constantly increasing and to keep low-cost service, IT service providers are aiming to investigate their energy efficiency. In addition to cost reduction, IT service providers are recently also obliged to reduce their CO_2 emissions under global and governmental pressures [2, 4, 6]. Yet, those facts draw the attention of both industry and academia and give a solid motivation to investigate the energy efficiency of data centers in addition to the ever-observed indicator ‘High Performance’.

In this paper, we will present a holistic view of the virtual machines placement problems, their challenges and the currently used solution methodologies in the literature, which target reducing energy power consumption and aim for sustainable IT resources management. More precisely, the second section comprises preliminaries of the problem and a view on different fields of research that heavily addressing the problem. In the third section, we will present a case study to stress on the necessary simplicity in analyzing the system requirements before designing solution strategies for reducing energy power consumption. Followed, the fourth section contains an implementation overview and the computational results of the conducted simulation study before closing the paper with a conclusion and further research directions.

2 Literature review

Virtual machines live migration is a recent topic in addition to some others, in which the allocation of resources in data centers is investigated to accomplish an efficient

energy consumption by migrating active virtual machines between available physical hosts [7]. A clear definition of the problem has been previously presented and widely discussed under Virtual Machine Placement Problem (VMP) in the literature [8, 9], where virtual machines must be allocated to functional physical hosts using a specific load distribution strategy. The allocation process is successfully conducted only when the required information of a virtual machine and its required computing resources is passed to the load distributor, which allocates it to a functional server, that possesses and can satisfy those requirements [9].

Many similar problems have been previously investigated in the literature and unfortunately proven to be NP-complete even in best scenarios as for instance the bin-packing problem [10] or the identical parallel machine scheduling problems [11]. The bin-packing problem is still, under the formal description, considered to be easier to tackle down than placing virtual machines on physical hosts. Those because a couple of additional constraints by the VMP must be taken into consideration in order to avoid undesired behaviors or instability in the system [9].

An independent stream of research deals explicitly with the live migration process of virtual machines, which focuses more precisely on the migration process itself and how it can be conducted with the minimum downtime of the service to minimize the impacts of migration processes on the system stability and on the signed Service Level Agreement (SLA). In this field, detailed analysis of the migration process in addition to many other implementation concepts is discussed. One of the prior works in this field was presented in [12]. The authors presented a framework and detailed implementation strategy for virtual machine live migration, which minimizes the downtime of the process. They described the migration strategy under six different stages, which might be classified to three main phases in terms of migrating the memory. First is the push phase, in which certain pages of the migrated virtual machine is iteratively pushed to the destination host. The second phase comprises the completion of the copy process of the old Virtual Machine (VM) to the destination host, stopping the old VM and starting the new VM on the new host. In the third phase, after executing the new VM, they make sure that all the required information of the migrated VM have been successfully copied and if not tolerate the missing information and pull them from the old source. Other local resources of the migrated VMs are also directly migrated to the new host to avoid IP forwarding mechanisms, which as they describe might have implication on the system stability. They reported impressive results with a downtime, which did not exceed 60 milliseconds.

The allocation or the placement of virtual machine to which physical host using different load distribution strategies is the oldest field of research that explicitly deals with the VMP problem. Many load distribution strategies have been discussed and presented in the past two decades, since the introduction of distributed computing systems [13–15]. Two main types of load distribution strategies can be distinguished, static load distributors and dynamic load distributors [16]. A well-known example for the static load distributor is the Round & Robin scheduling algorithm [17], which is also classified under load balancing strategies [18]. Static load distributors are often easy and intuitive to implement and possess the advantage of light execution time to take a decision for a new allocation. However, they are on the other hand not aware of

the current load of the different hosts and may lack on performance issues and cause drawbacks in many instances, since they do not observe the dynamic behavior of the system [16]. A dynamic load distributor constantly observes the status of the functional hosts in the system during the operational time to adapt the current best allocation for new machines based on a specific strategy, which can be set differently [16, 18].

For instance, in [19], the authors presented a central dynamic load distribution algorithm. They proposed a load balancing strategy to treat load distribution problem of web-servers application. They defined the incoming requests as a set of jobs that must be allocated to different localhosts or machines. Their strategy constantly observes the current load of the servers in terms of CPU and labels them as heavy, moderate, or high loaded. This information is exchanged with the master server that hosts the load balancer, which initiates migration policies based on those labels until reaching a balanced loaded system. The exchange process is conducted periodically using a 10 minutes interval. They, however, stressed at the end of their study on taking the memory capacity of servers into consideration in the future work.

Since the majority of the problems in this field are quite complex and classified under NP-complete class of problem, the research was more focused on finding and presenting sub-optimal solutions [15]. Consequently, heuristics were the dominant solution approaches to deal with the VMP problem in the past two decades with over 67 % adaptation of the total literature analysis, which has been concluded in the survey presented in [8]. Nevertheless, some attempts to adapt meta-heuristics to investigate load distribution strategies can be found [8]. For instance, [18] presented a solution approach for load balancing using genetic algorithms and presented an analysis on a small scale. The authors considered in their study and computational analysis up to three physical hosts and 48 virtual machines. They formulated a fitness function, which took the CPU and memory utilization into consideration. They compared their results to two other heuristics namely, Round & Robin and greedy algorithm.

Recently, a parallel stream of research in this field has been focused on studying and introducing load distribution strategies, which mainly target energy consumption and adopt it as an objective function to reduce operational costs [2, 5, 8, 20]. For instance, in [2], an energy-aware load distribution strategy has been presented to efficiently manage data centers. The authors presented a sophisticated architectural framework, in which different types of virtual machines with different SLAs are taken into consideration in the allocation decision-making process. A major part of their framework was the presented power model, which they use in their algorithm. They assumed that the CPU consumes the most power in a physical host. Based on that assumption they assigned an estimated fraction of power consumption for every new virtual machine in the system. The precise implementation of their allocation algorithm was based on the Best Fit Decreasing algorithm, which they modified to take decisions based on power consumption of CPU after allocating a machine to it. They reported an extensive analysis based on a simulation study of their work and recorded SLA violations. They compared their strategy with different non-power aware strategy and reported a decrease in energy consumption in comparison to all of them. In the next section, a case study is presented to demonstrate the role of load distribution strategies in reducing energy power consumption by IT service providers.

3 A case study

The major challenge in designing load management strategies lies in understanding the nature of the incoming workload patterns and their characteristics. Since the heterogeneity of the incoming workload patterns is considerably high, the presented solution approaches in the literature are either problem-specific or highly generic. Both types suffer major drawbacks in terms of applicability and designed objective function. In order to manage resources in data centers taking sustainability and energy consumption, predicting the incoming load is very valuable since the scaling process might take time in data centers due to the normal provisioning processes. This could cause either a degradation in the quality of service or an over-provisioning that is associated with additional energy consumption to meet the signed SLAs [21]. Many contributions in this field have been presented using different methodologies starting by traditional statistical methods [22]. Most recently, very sophisticated approaches as for instance machine learning and neural network [23] have been adopted to understand the different incoming workload patterns in data centers. The main goal of those attempts is to increase the efficiency of the automatic scaling in the cloud since it is one of the main advantages of cloud computing [22, 23].

Yet, with the introduction of cloud computing model, businesses are even more motivated to seize the advantages of scalable outsourced IT systems [24]. Those represent a significant proportion of the overall customers by IT service providers [24]. Although the prediction process of the required computational power is quite tedious, this fraction of customers normally generates anticipated request patterns with a stable workload behavior. Their workload behavior might be even predicted on a daily basis using a very naïve approach. The behavior of that customer fraction can be often related to the normal working hours in a company, which outsources its IT system and offers their employees desktop accesses. This implies, that we can even describe their Virtual Machines (VMs) active hours during a day using mathematical distributions. For instance, the employees of a German company should usually fulfill 40 working hours a week. If we assume, that their VMs are active during their working hours, we can roughly predict their required computational power on the real physical host to efficiently allocate them to reduce energy consumption.

3.1 Closed class interactive workload queuing network for IT system landscape

The behavior of the previously fraction of customers by IT service providers can be classified under the closed class interactive workload queuing network, which is presented in [25]. The model is initially introduced for classifying interactive workload patterns of database servers. This class has a certain behavior and characteristics according to the authors, which are summarized in the following:

- The number of customers specifies the workload intensity.
- Number of customers is bounded, known and can be used as a parameter.
- Throughput is calculated after solving that queuing network and based on the customer population of that class.

- After a request is processed, the customer sends the next request after a 'thinking time'.

In order to investigate and further argue the applicability of this concept for a sustainable management of IT resources, a case study is presented. The main motivation behind the case study is to support the previously discussed arguments regarding the used dimension for allocating the resources to the virtual machines. In addition, we want to shade a light on the advantages, which can be seized through using simple information that describe the workload behavior of big customers by IT service providers.

3.2 Problem formulation and objective function

Given a data center, which consists of a set of physical machines and a corresponding queue of clients, which demands a set of virtual machines to be deployed on those hosts. The virtual machine placement problem is investigated and solved based on many objective functions as for instances maximization of the system performance or the minimization of operational costs. The VMP might be formalized in the following:

- Let $H = \{h_1, \ldots, h_m\}$: be a set of m hosts.
- Let $V = \{v_1, \ldots, v_n\}$: be a set of n online virtual machines.
- Let $R = \{r_1, \ldots, r_o\}$: be a set of o resources required for each $v \in V$.
- Let $S = \{s_1, \ldots, s_m\}$:be the set of m values, which represent the shutdown hours of the hosts $H = \{h_1, \ldots, h_m\}$ during a time span T.
- Let $D_{i,y}$: be the required resource for $v_i \in V$ from resource type $y \in R$
- Let $C_{j,y}$: be the total capacity of $h_j \in H$ of the resource type $y \in R$

It is desired to allocate this set of VMs V on the hosts dynamically, in which the total shutdown hours of all servers over a time interval T is to be maximized as shown in equation 1. Those are to reduce total energy consumption taking into consideration the corresponding SLAs requirements and conditions.

$$Maximize\ total\ shutdown\ hours = \sum_{x=1}^{m} s_x : subject\ to\ (2) \qquad (1)$$

$$\forall\, y \in \{r_1, \ldots, r_o\} : \sum_{i=1}^{n} D_{i,y} < \sum_{j=1}^{m} C_{j,y} \qquad (2)$$

3.3 System description and specifications

The IT landscape of the considered system consists of eight homogeneous servers, which host five different types of virtualized system deployed in 290 VMs as presented in Table 1. The considered system provide SAP system access as a service for high educational institutions with different characteristics. In addition, they provide desktop access for research purposes. The main memory capacity of each server is 500 GB. Nevertheless, the bottleneck in the system is believed to be the main memory, since a fixed allocation of the main memory is maintained during the deployment of the VMs. In addition, the monitoring information, which has been extracted from the system, showed that the generated CPU workload by the VMs is relatively low. We described the online hours of the different VMs based on expert's interviews using mathematical distributions. For instance, a researcher works roughly 40 hours a week with an average of 8 hours a day. In addition, the periods, in which the VMs are active, are mostly during the day.

Table 1. The specifications of the deployed virtual machines in the considered IT landscape.

Virtual machine Type	*Main memory*	*Quantity*	*Online time distribution*	*Offline time distribution*
Research assistant VMs	4	30	Triangular (1, 6, 3)	Triangular (22, 30, 24)
Researcher VMs	8	30	Triangular (6, 14, 8)	Triangular (14, 18, 16)
SAP system access 1	10	90	Triangular (2, 8, 5)	Triangular (16, 22, 19)
SAP system access 2	12	40	Triangular (2, 8, 5)	Triangular (16, 22, 19)
SAP system access 3	14	100	Triangular (2, 8, 5)	Triangular (16, 22, 19)

3.4 Conceptual representation of the simulation model

The evaluation process of new or existed load distribution strategies is normally conducted in an emulated environment of the real system. Those to avoid instability and other implications (e.g. network traffic) in the real system during operational time by using new distribution strategy, which might cause violations in the SLA [20]. For instance, simulation is a very powerful technique, which is widely used to investigate different aspects of IT system landscapes [26] and might be combined with various optimization strategies [27]. Recently, many new simulation packages have been developed to model and investigate cloud computing environments [28].

In Figure 1, the conceptual representation of the simulation model is presented, which is derived from the basis of the previously discussed closed class interactive workload queuing network, to model IT system landscape. An initial number of virtual machines is created in the closed queuing network. Different types of VMs can be taken

into consideration. Those VMs must be dynamically allocated during the operational time to *n* number of active servers. According to this concept, the virtual machines do not leave the system. Instead, the virtual machines have two main states in the system, an active state, when the VMs are online, and a passive state, when the VMs are offline. The passive state is known in the original model as the thinking time. During the thinking time, the user does not generate any workload in the system, which corresponds to the concept of offline hours by a VM. In the next section, will discuss the implementation of the simulation model demonstrate the obtained computational results.

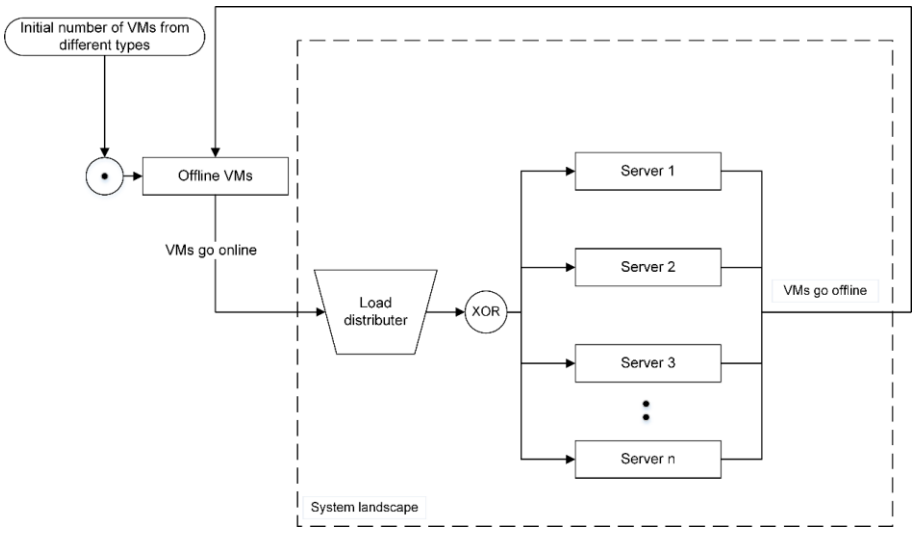

Figure 1: Conceptual representation of the simulation model based on the closed class interactive workload queuing network for IT landscape.

4 Implementation and computational results

Since new load distribution strategies are not supposed to be, without previous evaluation, directly used during the operational time, a prototypical implementation of the presented concept has been developed in the form of a simulation study. The IT landscape of the considered system and its specification have been mapped to the simulation model. The simulation model has been built in ExtendSim simulation package. Discrete event simulation approach has been adapted to build the simulation model. We evaluated two different scenarios using different load distribution strategies to identify the most suitable strategy for reducing energy power consumption.

4.1 The evaluated load distribution algorithms

We evaluated load balancing and load concentration algorithms on the modeled system with the given specifications. A control policy based on lower- and upper-threshold is

designed to conduct live migrations on the virtual machines. A 25 % local lower threshold on the physical hosts has been evaluated to initiate migrations in the system. Moreover, the control policy is executed whenever a virtual machine goes offline to decided whether to trigger a live migration or not. The migration policy is conducted only to switch off an active server if this desired migrated load can be hosted on the other active servers in the system. However, a 20 % global upper threshold of the RAM capacity is evaluated to automatically scale up the system capacity through activating hibernated physical servers.

The load concentration algorithm

Since the bottleneck of the system, described in the use case, is the main memory, the algorithm is designed to allocate VMs based on the main memory utilization of the servers. The load concentration algorithm is based on the best-fit decreasing algorithm, which has been modified to allocate the VMs based the main memory. Moreover, the algorithm maintains a list of the available capacity of all active servers sorted in an increasing order. It is a very straightforward algorithm, which concentrates the load on the currently most loaded server. This is to reduce the number of initiated migration policies since the least loaded server maintains its status and does not receive any new VMs if not needed.

The load balancing algorithm

The load balancing algorithm is similarly designed to the load concertation one. It is also based on the best-fit decreasing algorithm, which has been modified to allocate VMs based on the main memory utilization of the servers. However, this algorithm maintains a list of the available capacity of all servers sorted in decreasing order. The algorithm aims to sustain a balance in the workload between servers by allocating VMs to the least loaded server, which is the first server in the list. The migration policy of this strategy is based on the threshold principle.

4.2 Evaluation and computational results

The demonstrated distributions in Table 1, has been used to model the processes of the virtual machines in the system. Two algorithms are tested in combination with migration and control policies. Eight servers have been taken into consideration with 500 GB main memory capacity for each. The simulation has been set to consider a time interval of 120 hours, which correspond to five days. For each simulated scenario, 200 replications are recorded to ensure the quality of the obtained results and eliminate the byes from the system. The average total shutdown hours of all servers, the average number of initiated migration policies and the average of the total migrated VMs have been closely observed over the conducted simulation runs. A 95 % confidence interval has been applied on all observed measurements to observe the possible deviation and

obtain the margin error. The computational results of the evaluated scenarios are presented in Figure 2.

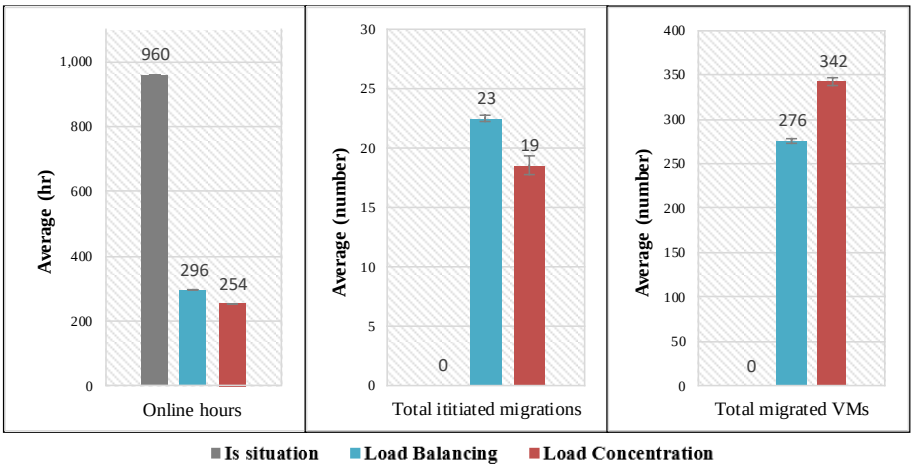

Figure 2: The computational results of the simulation model.

With a simple calculation, the total possible online hours of all servers over the simulation time is equal to the number of servers multiplied by the number of simulated hours of five days, which is 960 online hours. Both load distribution strategies manage the load of the described system using roughly between 25% and 30 % of the total available online hours of all servers. The load concentration strategy outperforms the load balancing one in terms of maximizing the total shutdown hours. However, the load balancing algorithm reports a significant outperformance in terms of the number of migrated VMs during the considered time span. Although as expected, the number of initiated migration policies by the load concentration algorithm is slightly smaller, the number of migrated VMs is considerably higher than the one, reported by the load balancing algorithm.

5 Conclusion and future work

The analysis showed that the load concentration algorithm performs the best to reduce the energy consumption of the considered IT landscape. Nevertheless, we recommend applying load balancing algorithm with a global lower threshold to maximize total shutdown hours since this algorithm maintains a higher stability in the system in terms of the number of migrated virtual machines. Nowadays, a considerable proportion of cloud customers are small or middle size companies, which outsource their entire IT requirements based on the concept of cloud computing. Those, reflect a high potential in managing their system efficiently by an IT service provider through seizing the advantage of their predicted generated load behavior. Future works might spot a light on a hybrid load distribution strategy for managing IT system landscapes. Moreover,

investigating a large-scale use case, in which several independent closed IT system landscapes are taken into consideration, would generates a federated or autonomic management of load distribution strategies in data centers. Let us assume that a definite number of closed classes can be indicated in a data center. Accordingly, we can classify them into classes and separate their load distribution algorithms. In other words, the load distribution strategies can be implemented and offered as a service to manage the workload of a data center to satisfy various SLAs requirements. Based on the requirements of various systems, a distribution algorithm with the required allocation dimension (e.g. RAM, CPU or a combination) might be used to manage their generated workload in the data center. This federation of managing the data center might allow IT service providers to manage the load distribution more efficiently, by seizing the advantage of separating unpredicted workload system (e.g. web application of online shop) from the predicted ones. In addition, solving smaller independent sub-problems and managing resource allocation in data centers based on autonomic strategy might contribute in gaining major advantages over highly generic solution for a sustainable IT resources management.

References

1. Mell, P., Grance, T.: The NIST definition of cloud computing. NIST special publication 800, 1–3 (2011)
2. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Generation Computer Systems 28, 755–768 (2012)
3. Li, K., Wu, J., Blaisse, A.: Elasticity-aware virtual machine placement for cloud datacenters. In: Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on, pp. 99–107 (2013)
4. Koomey, J.: Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times 9 (2011)
5. Splieth, M., Kramer, F., Turowski, K.: Classification of Techniques for Energy Efficient Load Distribution Algorithms in Clouds-A Systematic Literature Review. In: EnviroInfo, pp. 605–612 (2014)
6. Hawkins, A.: Unused Servers Survey Results Analysis (2010)
7. Orgerie, A.-C., Assuncao, M.D.d., Lefevre, L.: A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Computing Surveys (CSUR) 46, 47 (2014)
8. Lopez-Pires, F., Baran, B.: Virtual machine placement literature review. arXiv preprint arXiv:1506.01509 (2015)
9. Hyser, C., Mckee, B., Gardner, R., Watson, B.J.: Autonomic virtual machine placement in the data center. Hewlett Packard Laboratories, Tech. Rep. HPL-2007-189, 1–10 (2007)
10. Skiena, S.S.: The algorithm design manual: Text. Springer Science & Business Media (1998)
11. Pinedo, M.: Scheduling. Theory, algorithms, and systems. Springer, New York (2012)

12. Clark, C., Fraser, K., Hand, S., Hansen, J.G., Jul, E., Limpach, C., Pratt, I., Warfield, A.: Live Migration of Virtual Machines. In: Proceedings of the 2Nd Conference on Symposium on Networked Systems Design & Implementation - Volume 2, pp. 273–286. USENIX Association, Berkeley, CA, USA (2005)
13. Pinheiro, E., Bianchini, R., Carrera, E.V., Heath, T.: Load balancing and unbalancing for power and performance in cluster-based systems. In: Workshop on compilers and operating systems for low power, 180, pp. 182–195 (2001)
14. Williams, R.D.: Performance of dynamic load balancing algorithms for unstructured mesh calculations. Concurrency: Practice and experience 3, 457–481 (1991)
15. Cybenko, G.: Dynamic load balancing for distributed memory multiprocessors. Journal of Parallel and distributed Computing 7, 279–301 (1989)
16. Mohapatra, S., Rekha, K.S., Mohanty, S.: A comparison of four popular heuristics for load balancing of virtual machines in cloud computing. International Journal of Computer Applications 68 (2013)
17. Shreedhar, M., Varghese, G.: Efficient fair queuing using deficit round-robin. IEEE/ACM Transactions on networking 4, 375–385 (1996)
18. Chandrasekaran, K., Divakarla, U.: Load Balancing of Virtual Machine Resources in Cloud Using Genetic Algorithm. In: (2013)
19. Bhadani, A., Chaudhary, S.: Performance evaluation of web servers using central load balancing policy over virtual machines on cloud. In: Proceedings of the Third Annual ACM Bangalore Conference, p. 16 (2010)
20. Splieth, M., Bosse, S., Schulz, C., Turowski, K.: Analyzing the Effects of Load Distribution Algorithms on Energy Consumption of Servers in Cloud Data Centers. In: 12th International Conference on Wirtschaftsinformatik (2014)
21. Xiong, K., Perros, H.: Service performance and analysis in cloud computing. In: 2009 Congress on Services-I, pp. 693–700 (2009)
22. Yoon, M.S., Kamal, A.E., Zhu, Z.: Request Prediction in Cloud with a Cyclic Window Learning Algorithm. arXiv preprint arXiv:1507.02372 (2015)
23. Islam, S., Keung, J., Lee, K., Liu, A.: Empirical prediction models for adaptive resource provisioning in the cloud. Future Generation Computer Systems 28, 155–162 (2012)
24. Motahari-Nezhad, H.R., Stephenson, B., Singhal, S.: Outsourcing business to cloud computing services: Opportunities and challenges. IEEE Internet Computing 10, 1–17 (2009)
25. Menasce, D.A., Almeida, V.A.F., Dowdy, L.W., Dowdy, L.: Performance by design: computer capacity planning by example. Prentice Hall Professional (2004)
26. Núñez, A., Vázquez-Poletti, J.L., Caminero, A.C., Castañé, G.G., Carretero, J., Llorente, I.M.: iCanCloud: A Flexible and Scalable Cloud Infrastructure Simulator. Journal of Grid Computing 10, 185–209 (2012)
27. Bosse, S., Splieth, M., Turowski, K.: Multi-objective optimization of IT service availability and costs. Reliability Engineering & System Safety 147, 142–155 (2016)
28. Splieth, M., Bosse, S., Turowski, K.: Analysis Of Simulation Tools For Determining The Energy Consumption Of Data Centers For Cloud Cmputing. In: Bruzzone, A., Felice, F., Massei, M., Merkuryev, Y., Solis, A., Zacharewicz, G. (eds.) Proceedings of the 13th International Conference on Modeling and Applied Simulation. Curran Associates, Inc., France (2014)