

Toward an Autonomic and Adaptive Load Management Strategy for Reducing Energy Consumption under Performance Constraints in Data Centers

Abdulrahman Nahhas^a, Sascha Bosse^b, Matthias Pohl^c and Klaus Turowski

Very Large Business Applications Lab, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Germany

Keywords: IT Resources Management, Adaptive Load Distribution Strategies, Virtual Machines Live Migration, Energy-aware Virtual Machines Allocation, Heuristic and Metaheuristic Optimization, Data Center Management.


Abstract: The future vision of IT-industry is shifting toward a utility-based offering of computing power using the concepts of pay-per-use. However, the elasticity and scalability characteristics of cloud computing massively increased the complexity of IT-system landscapes, since market leaders extensively expanding their IT-infrastructure. Accordingly, the carbon-footprint of data centers operations is estimated to be the fastest growing footprint among different IT fields. The majority of contribution in the examined literature that address IT resources management in data centers exhibits either a specific or a generic nature. The specific solutions are designed to solve specific problems, but yet neglecting the dynamic nature of IT-systems. The design of generic solutions usually overlooks many details of the investigated problems that have an impact on the possible optimization potential. One can argue that an optimized combination of different algorithms used during a specified time span would outperform a single specific or generic algorithm for the management of IT resources in data centers. Therefore, a conceptual design for an autonomic and adaptive load management strategy is presented to investigate the aforementioned hypothesis. Our initial experimental results showed considerable improvement when multiple algorithms are used for the allocation of virtual machines.


1 INTRODUCTION


Virtualization strategies have changed the traditional design and deployment of IT-system landscapes. The future vision of IT-industry is shifting toward a utility-based offering of computing power using the concepts of cloud computing. Therefore, market leaders are massively expanding their IT-system landscapes (Kushida *et al.*, 2011), in which the optimization of the IT-system design and engineering is not significantly important for decision makers to announce investments worth millions of euro for new IT-infrastructure. However, the elasticity and scalability features of the cloud computing model have a major impact on the complexity of IT-system landscapes, since the incoming workload becomes much harder to predict. Consequently, the massive expansions of IT-system landscapes in addition to the

mentioned characteristics of cloud computing radically complicated the management process of those landscapes.

The energy costs will keep increasing, which poses a necessity for IT-service provider to investigate the efficiency of their operations to reduce costs while holding their Service Level Agreements (SLA). The efficiency of utilizing IT-resources becomes a market competitive advantage for IT-service provider to offer reliable but yet sustainable IT-services with reasonable costs in the market. The main fraction of costs is encountered through the energy consumption of physical servers, which is estimated to reach up to 50 % of the overall costs. In addition, statistical analysis on the worldwide energy consumption triggered an alarm on a governmental level since numbers suggest a total growth of roughly hundred percent reported by data center industry

^a  <https://orcid.org/0000-0002-1019-3569>

^b  <https://orcid.org/0000-0002-2490-363X>

^c  <https://orcid.org/0000-0002-6241-7675>

between 2005 and 2010 (Kooimey, 2011). Obviously, the associated CO₂ emissions of data center's operations reported, accordingly, tremendous growth and estimated to be the fastest growing carbon-footprint among different IT fields (Avgerinou *et al.*, 2017). On a European Union (EU) level, some initiatives, research, and further regulations have been introduced to suppress the impact of data center CO₂ footprint as, for instance, the EU Data Center Code of Conduct (Avgerinou *et al.*, 2017). Those facts motivated data centers operators to revision the management strategies of data centers to achieve a higher level of sustainability in service offering and management.

In this research, we will present an overview on the current advances of load management strategies targeting sustainable management of IT resources in data centers. The second section is dedicated to shed a light on the usual formulation of static and dynamic Virtual Machines (VMs) placement problems. Based on the initial findings, we present our intermediate analysis from our systemic literature analysis. We discuss the adopted solution approaches from an algorithmic point of view to present our hypothesis and research question. In the third section, a conceptual design for an autonomic and adaptive load management strategy is presented. The fourth section is dedicated to present the initial computational results to answer the posed research before closing the paper with a conclusion.

2 STATE OF THE ART AND LITERATURE ANALYSIS

Many efforts and investigations have been dedicated in the last two decades to propose efficient but yet specific solutions for data center management. The majority of the static virtual machines placement or virtual machines consolidation problems are formulated in different forms of bin-packing problems (Lopez-Pires and Baran, 2015). The simplest form is the single dimension bin-packing problem taking the CPU as the main resource to allocate the virtual machines. The goal is eventually to place the existing virtual machines modelled as items into the minimum number of active physical hosts modelled as bins to reduce the overall energy consumption. Many similar problems have been intensively addressed in the literature in the fields of scheduling and operations research as for instance, the identical and non-identical parallel machines scheduling problems and different forms of bin-

packing problems (Pinedo, 2012; Skiena, 1998). Unfortunately, the majority of those problems have been proven to be NP-Hard. In addition, the complexity of a considered problem is further increased when it is formulated in form of multi-dimensional bin packing problem. In such more realistic problem formulation, three recourses dimensions can be taken into consideration as for instance, CPU, memory and storage.

Therefore, the majority of the research conducted on the virtual machines placement problems is inspired by heuristic approaches. They are usually adopted when the solution space of a problem cannot be investigated entirely with the current computational power in polynomial time. More profoundly, heuristic approaches comprise two main categories: constructive and improvement approaches. The constructive approaches are simple straightforward algorithms, in which the decision for allocation is taken instantly after conducting some calculations without searching in the solution space of the problem. They are intuitive to implement and exhibit a light execution time to take decision for a new allocation. However, they are not robust against major modifications in the problem formulation. In essence, if the underlying infrastructure or the incoming workload patterns witness a major change, their performance usually massively degrades and their internal design has to be adjusted accordingly (Keller *et al.*, 2012).

Therefore, IT- research has been for decades relying on improvement and metaheuristic approaches for solving static virtual machine consolidation problems. Improvement heuristics are conceptually more sophisticated heuristic procedures in comparison to the constructive ones since the construction of a solution is the first step in their internal functionality. Thereafter, based on a solution, an improvement heuristic seeks to conduct single or several changes on the constructed allocation to find a so-called neighbour solution, which hopefully yields to a better investigated objective function. The modification process is then iteratively conducted until some breaking criterion is met. Finally, the metaheuristic approaches are the most powerful optimization techniques that fall under heuristic procedures. The majority of them are inspired by some natural phenomena, as for instance, Genetic Algorithms (evolution theory) (Holland, 1992) or Simulated Annealing (annealing process of metals) (Kirkpatrick *et al.*, 1983). They are fundamentally based on an improvement heuristic and an overall control strategy that attempt to guide the

improvement procedure to achieve better optimization results and avoid false or local optima.

However, the adoption of those optimization techniques is associated with significantly higher computational effort to find good solutions in comparison to constructive approaches. Therefore, their adoption is strictly subject to whether the allocation decision needs to be taken instantly or not. In the static virtual machine consolidation problem case, the required computational effort to find a very good to a near optimal solution does not have to be necessarily instant since the migration of the virtual machines is conducted anyways in an offline mode.

In addition to the rapid evolvement of virtualization strategies, the introduction of virtual machines live migration algorithms shifted the focus of academia from the classical static virtual machines consolidation problems to the so-called dynamic Virtual Machines Placement problems (VMP). Dynamic virtual machine placement implies that virtual machines are subject to reallocation processes during operational time based on the dynamic state of the system to meet various goals, as for instance, to reduce energy consumption. The virtual machines live migration algorithms address the migration process of virtual machines during the operational time, in which the goal is to migrate a virtual machine from an active physical host to another one with the minimum downtime (Clark *et al.*, 2005; Jin *et al.*, 2014). Thus, to reduce the impact of the migration process on the associated hosted services in order to avoid violations in the signed SLA while reducing the overall energy consumption. This major advance in virtualization strategies led to the introduction of a new research stream under the term “energy-aware”. In the past decade, many algorithms have been presented to schedule virtual machines or tasks taking into consideration the increase in the energy consumption of the underlying infrastructure. In addition, the popularity of metaheuristic approaches for virtual machines allocation is significantly decreased with major domination of heuristics approaches. Obviously, the reason can be traced back to the relatively high required computational effort of them to find suitable allocation. Based on our initial analysis on the prominent publications on science direct database the majority of the found articles are presenting energy-aware solutions such as the contributions of (Zheng and Cai, 2011; Goiri *et al.*, 2012; Bodenstein *et al.*, 2012; Beloglazov *et al.*, 2012; Luo *et al.*, 2013; Zhang *et al.*, 2014; Tesfatsion *et al.*, 2014; Khani *et al.*, 2015; Dupont *et al.*, 2015; Kumar and Raghunathan, 2016; Carli *et al.*, 2016; Vafamehr and Khodayar, 2018; Marotta *et al.*, 2018;

Malekloo *et al.*, 2018; Kaur and Chana, 2018; Han *et al.*, 2018). Energy-aware heuristics are specially designed algorithms to reduce energy power in data centers and usually based on a core power model that highly determines their behavior. The majority of the found articles are based on constructive heuristic procedures since the light execution time is of major importance for solutions with live migration capabilities. For instance, in (Beloglazov *et al.*, 2012; Dupont *et al.*, 2015; Han *et al.*, 2018; Zhang *et al.*, 2014) the core functionality of the presented heuristics is based on the designed or adopted power model.

However, IT industry requires solutions that are able to adapt to the dynamic nature of those systems with minimal human intervention. Therefore, the complexity of our current and future IT-systems requires a deep analysis of the current understanding of artificial intelligence techniques and its advances for automation proposes. Therefore, based on our initial literature analysis, we identified two main research streams that explicitly deal with the virtual machines placement problem based on some machine learning approaches to reduce energy consumption in data centers. In the first stream, adaptive approaches are presented such the contributions of (Jeyarani *et al.*, 2012; Xu *et al.*, 2012; Vitali *et al.*, 2015; Suresh and Sakthivel, 2017; Yoon *et al.*, 2017; Zhou *et al.*, 2018; Kumar and Singh, 2018). Adaptive solutions are algorithms with monitoring capabilities that are designed to react or adapt to specific scenarios such upper and lower threshold of server’s workload or statistical analysis on workload to rely on some predictions (Yoon *et al.*, 2017). The internal design of the presented solutions is definitely more sophisticated than the energy-aware solutions and the majority of them have been presented in form of frameworks. In essence, in the adaptive solutions, the algorithm is a component that relays on some prediction model to derive predictions for the upper and lower threshold of servers to take the allocation decision as in (Jeyarani *et al.*, 2012; Kumar and Singh, 2018; Suresh and Sakthivel, 2017; Vitali *et al.*, 2015; Yoon *et al.*, 2017; Zhou *et al.*, 2018).

The second stream presents autonomic frameworks as for instance the contributions of (Wang *et al.*, 2008; Xu *et al.*, 2012; Tchana *et al.*, 2013; Amoretti *et al.*, 2013; Delaval *et al.*, 2015). Autonomic strategies are self-organizing strategies that exhibit sophisticated features usually targeting the management of landscape on an application level as for instance, application scalability (Tchana *et al.*, 2013; Wang *et al.*, 2008; Delaval *et al.*, 2015). Based on the conducted analysis, the majority of

contributions in the examined literature exhibit either a specific or a generic nature. The specific solutions are designed to solve specific problems, but yet neglecting the dynamic nature of IT-systems especially in a cloud-computing context. The design of generic solutions usually overlooks many details of the investigated problems that have an impact on the desired optimization potential and thus, do not achieve the possible optimization potential. Therefore, we aim to answer the following research question: Will a combination of heuristic and metaheuristic approaches to present a hybrid framework for the management of data center operation overcome the aforementioned drawbacks in the analyzed literature?

The question is based on the argument that an optimized combination of different algorithms used during a specified time span would outperform a single specific or generic algorithm for the management of IT resources in data centers. To exactly know how the combination should be built, we need to rely on some overall optimization mechanisms, as for instance, a metaheuristic approach. The main idea is to exploit the light execution time of constructive approaches to take instant decision for allocation and the robustness of metaheuristic approaches to achieve a higher optimization potential. In the course of the next two sections, we present a conceptual design of an adaptive and autonomic concept for the management of data centre operations based on multiple algorithms to answer the research question and validate the aforementioned hypothesis.

3 ADAPTIVE AND AUTONOMIC LOAD DISTRIBUTION STRATEGY

The concept under design is presented in Figure 1. The framework consists of three main components: the workload Monitoring and Prediction component (MP), the Adaptive component (AD) and the Artificial Intelligence component (AI). The MP component is designed to deliver likely future workload distributions of the considered Virtual Machine (VMs) types. Based on the analyzed literature, one can rely on statistical analysis or machine learning approaches on the workload demand to predict the incoming workload for a specific time span (Kumar and Singh, 2018). In some studies, it is even suggested to conduct statistical analysis on the power consumption requirements on

an application level to derive power consumption profiles of applications. For instance, Bartalos *et al.* (2016) presented an aggregated model to predict the power demand of an application running on specific servers using multiple linear regression models. It is of interest to study the behavior of the optimization model if one combines both prediction approaches to derive workload profiles as well as energy power profiles.

The AD component contains an optimization and evaluation models. In our prototypical analysis and implementation, we relied on Genetic Algorithms (GA) to design the optimization model, which is dedicated for finding the best combination of heuristics that should be used for load management depending on the system state over time. The optimization model might be further fed with different algorithms, performance models, operational constraints and finally different sensitive parameters that are collected through feedback loops from the AI component. As for operational constraints, different forms of Service Level Agreements (SLAs) can be modeled to suppress the impact of live migrations on the associated possible penalties. Sensitive parameters include performance-based and workload-based measurements, as for instance, the physical server's upper and lower thresholds and the global threshold of the system. Such measurements have a major impact on the design and the functionality of the scalability mechanisms of the systems. The evaluation model might be based on a simulation model.

The AD component provides a solution that contains a combination of different algorithms to be used during a defined time span for load management in addition to a set of sensitive parameters (e.g. Threshold of servers workload) to control the migration policies. The AI component is dedicated to learn from the optimization results, pass the solution to the underlined infrastructure and provide a feedback loop to continuously adjust the performance of the AD component to achieve better results in the next optimization interval. The goal of the feedback loops is mainly to reduce the deviation of the inquired predictions on the workload and other measurements from the actual ones and systematically achieving a higher accuracy of the optimization model. Fuzzy sets have been, for instance, applied to address different problems, especially, in the field of supply chain management (Ganga and Carpinetti, 2011). They are a powerful approach to model uncertainties of some phenomena and incorporating expert's knowledge. Thus, they can be adapted to model the highly dynamic behavior of IT landscapes and its associated

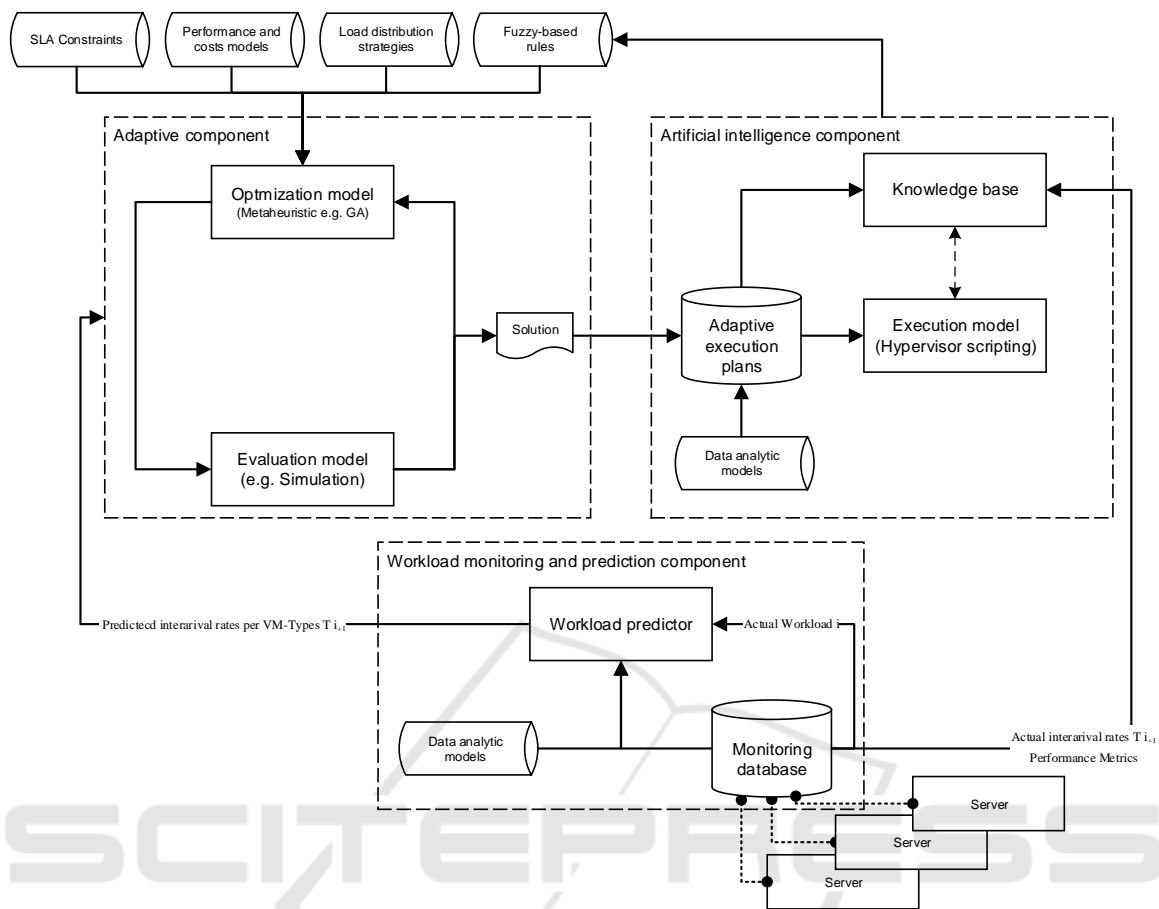


Figure 1: A conceptual model for an autonomic and adaptive load management strategy.

uncertainties. The overall goal is to detach the optimization component as soon as the AI component is trained and ready to work without the AD component. This goal can be systemically achieved through collecting data on the applied solutions and their deviation from the real data to derive measurements and sensitive parameters and apply data analytics approaches to extract knowledge. The obtained knowledge will be then further reflected in forms of rules and actions that must be applied to react to different phenomena.

4 INITIAL ANALYSIS ON THE PRESENTED CONCEPT

The initial analysis is dedicated to investigate the validity of the aforementioned hypothesis and answer the research question. Therefore, we relied on collected information through interviews with experts to mimic the MP component to derived different workload distribution for different VMs types of a

real system. In addition, we did not extend our experiment to investigate the role of the AI component since the research question is profoundly based on the functionality of the adaptive component. In the course of the next section, we present a brief problem formulation to investigate the functionality of the adaptive component before presenting our initial findings.

We relied on a simple problem formulation to draw some conclusions on whether the concept achieves the desired optimization potential or not (Nahhas *et al.*, 2018). One can assume that the optimization potential tends to increase with the increase in the complexity of a considered problem. The adaptive component is designed to optimize the functionality of the system, in which a simplified set of algorithms (load -concertation and -balancing) and sensitive parameters (Thresholds of physical servers) are passed to the optimization model. In our analysis, the optimization model is based on Genetic Algorithms (GA), while a simulation model is built to evaluate the fitness of the solution candidates. In this

initial analysis, we set the optimization model to investigate whether we need to change our allocation algorithm every hour. This implies that a solution candidate in the population of the GA comprises 24 integer values that represent the codes of the modelled allocation algorithms every hour and thresholds of physical servers.

We simulated five days of operations and relied on the expert’s interviews to derive mathematical distribution that describes the behavior of the considered virtual machines as shown in Table 1. The IT landscape of the considered system consists of eight homogeneous servers, which host five different types of virtualized systems deployed in 290 VMs. The capacity of the main memory of the servers is 500 GB. Unlike many problem formulations in the literature, the bottleneck of the considered system is not the CPU capacity but rather the main memory, since the majority of the offered virtual machines servers as desktops. This implies that sharing the main memory is not allowed. We formulated the problem to take into consideration the number of migrated virtual machines as well as the total number of online hours of all servers during a time span.

Table 1: Descriptive information of the virtual machines in the considered IT landscape.

VMs Type	Main memory	Online time	Offline time
Assistant VMs	4	Triangular [1, 6, 3]	Triangular [22, 30, 24]
Researcher VMs	8	Triangular [6, 14, 8]	Triangular [14, 18, 16]
SAP system access 1	10	Triangular [2, 8, 5]	Triangular [16, 22, 19]
SAP system access 2	12	Triangular [2, 8, 5]	Triangular [16, 22, 19]
SAP system access 3	14	Triangular [2, 8, 5]	Triangular [16, 22, 19]

Given a data centre, that consists of a set of physical machines, which are serving customer requests to deploy various types of virtual machines. The problem under investigation might be formalized in the following:

- Let $P = \{p_1, \dots, p_m\}$: be a set of m physical machines.
- Let $V = \{v_1, \dots, v_n\}$: be a set of n online virtual machines.
- Let $R = \{r_1, \dots, r_o\}$: be a set of o resources required for each $v \in V$.
- 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 $y \in R$.
- Let $A (A \in \{1, 2\})$: denote the codes that describe the algorithms that can be used for allocation of virtual machines.
- Let $S = \{s_1, \dots, s_m\}$: be the set of m values, which represent the online hours of the physical machines $P = \{p_1, \dots, p_m\}$ during a time span T .
- Let $M \in \mathbb{R}_+$ denote the number of migrated virtual machines over the time interval T .

Let \mathbb{H} denote the set of all possible combinations of the considered set of algorithms A during a defined time interval T . It is desired to find the combination of the algorithms $H \in \mathbb{H}$ to allocate the set of VMs V on the hosts dynamically. This combination is then subject to the minimization of γ_1 refers to the total online hours of all servers and minimization of γ_2 refers to the total number of migrated virtual machines over a time interval T as shown in equation (1). Those are to reduce total energy consumption taking into consideration the impact of live migration on the performance of the system in a simple formulation.

$$\gamma_1(H) = \sum_{x=1}^m S_x, \gamma_2(H) = M: \text{subject to (2)} \quad (1)$$

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

For solving the problem, we adopted a weighted-sum approach to formulate the objective function in formula (3) to obtain formula (4).

$$\min Z(H) \Leftrightarrow \min \gamma_1(H) \wedge \min \gamma_2(H) \quad (3)$$

$$\begin{aligned} \arg \min_{H \in \mathbb{H}} Z(H) &= W_1 \cdot \gamma_1 + W_2 \cdot \gamma_2 : \\ &\forall (W_1 + W_2 = 1) \end{aligned} \quad (4)$$

The simulation has been set to consider a time interval of 120 hours, which correspond to five days of operations. For the hybrid approach, 10 to 20 replications were recorded during the optimization before drawing any conclusion on the fitness of a solution candidate. Finally, after acquiring the solution, 200 replications are recorded to ensure the quality of the obtained results and eliminate the bias from the system for each simulated scenario. A 95 % confidence interval has been applied to all observed measurements to observe the possible deviation and obtain the margin of error. The results showed that the

hybrid approach significantly outperforms both algorithms in terms of minimizing the number of migrated virtual machines over the optimization interval. With a slight deviation from the load concentration algorithm, nearly the same performance in terms of minimizing the total online hours of physical servers is observed. The computational results of the experiments are presented in Figure 2. The 960 hours refers to the total online hours over all servers in the considered time interval. The obtained margin of error on the collected results in terms of the total migrated virtual machine ranged between (± 2.07 , ± 5.56). While more stable results are obtained in terms of the total initiated migrations that ranged between (± 0.21 , ± 0.67) and the total online hours (± 4.29 , ± 1.82).

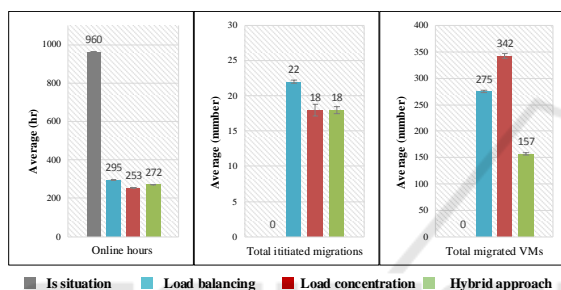


Figure 2: Experimental result on the presented hypothesis.

5 CONCLUSION AND FUTURE WORK

Our future work will be concentrated on finalizing a systematic literature analysis and further presenting taxonomy for the virtual machine live migration problems. Our preliminary analysis in a small use-case showed that the framework can achieve considerable improvements in minimizing the objective values. In addition, we are designing large-scale experiments based on collecting different information on the operational procedures of IT-service providers. Moreover, we are expecting to achieve a higher optimization potential with the increase of the problem complexity since the performance of the constructive approaches usually reasonable for solving simple problems. In the problem formulation, we addressed only the number of migrated virtual machine as an operational constraint, which might have an impact on the service level agreement. Therefore, in the final experimental analysis, we aim to address operational constraints by IT service provider more profoundly. In addition, in the formulated objective function we aim to address

not only the minimization of the total online hours but also different power states of server based on different workload levels to reduce energy consumption.

REFERENCES

- Amoretti, M., Zanichelli, F. and Conte, G. (2013), "Efficient autonomic cloud computing using online discrete event simulation", *Journal of Parallel and Distributed Computing*, Vol. 73 No. 6, pp. 767–776.
- Avgerinou, M., Bertoldi, P. and Castellazzi, L. (2017), "Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency", *Energies*, Vol. 10 No. 10, p. 1470.
- Bartalos, P., Wei, Y., Blake, M. B., Damgacioglu, H., Saleh, I. and Celik, N. (2016), "Modeling energy-aware web services and application", *Journal of Network and Computer Applications*, Vol. 67, pp. 86–98.
- Beloglazov, A., Abawajy, J. and Buyya, R. (2012), "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", *Future Generation Computer Systems*, Vol. 28 No. 5, pp. 755–768.
- Bodenstein, C., Schryen, G. and Neumann, D. (2012), "Energy-aware workload management models for operation cost reduction in data centers", *European Journal of Operational Research*, Vol. 222 No. 1, pp. 157–167.
- Carli, T., Henriot, S., Cohen, J. and Tomasik, J. (2016), "A packing problem approach to energy-aware load distribution in Clouds", *Sustainable Computing: Informatics and Systems*, Vol. 9, pp. 20–32.
- Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, C., Pratt, I. and Warfield, A. (2005), "Live Migration of Virtual Machines", in *Proceedings of the 2Nd Conference on Symposium on Networked Systems Design & Implementation - Volume 2*, USENIX Association, Berkeley, CA, USA, pp. 273–286.
- Delaval, G., Gueye, S. M. -K. and Rutten, É. (2015), "Distributed Execution of Modular Discrete Controllers for Data Center Management", *IFAC-PapersOnLine*, Vol. 48 No. 7, pp. 139–146.
- Dupont, C., Hermenier, F., Schulze, T., Basmaadjian, R., Somov, A. and Giuliani, G. (2015), "Plug4Green: A flexible energy-aware VM manager to fit data centre particularities", *Ad Hoc Networks*, Vol. 25, pp. 505–519.
- Ganga, G. M. D. and Carpinetti, L. C. R. (2011), "A fuzzy logic approach to supply chain performance management", *International Journal of Production Economics*, Vol. 134 No. 1, pp. 177–187.
- Goiri, Í., Berral, J. L., Fitó, J. O., Julià, F., Nou, R., Guitart, J., Gavalda, R. and Torres, J. (2012), "Energy-efficient and multifaceted resource management for profit-driven virtualized data centers", *Future Generation Computer Systems*, Vol. 28 No. 5, pp. 718–731.
- Han, G., Que, W., Jia, G. and Zhang, W. (2018), "Resource-utilization-aware energy efficient server consolidation

- algorithm for green computing in IIOT”, *Journal of Network and Computer Applications*, Vol. 103, pp. 205–214.
- Holland, J. H. (1992), *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press.
- Jeyarani, R., Nagaveni, N. and Vasanth Ram, R. (2012), “Design and implementation of adaptive power-aware virtual machine provisioner (APA-VMP) using swarm intelligence”, *Future Generation Computer Systems*, Vol. 28 No. 5, pp. 811–821.
- Jin, H., Deng, L., Wu, S., Shi, X., Chen, H. and Pan, X. (2014), “MECOM. Live migration of virtual machines by adaptively compressing memory pages”, *Future Generation Computer Systems*, Vol. 38, pp. 23–35.
- Kaur, T. and Chana, I. (2018), “GreenSched: An intelligent energy aware scheduling for deadline-and-budget constrained cloud tasks”, *Simulation Modelling Practice and Theory*, Vol. 82, pp. 55–83.
- Keller, G., Tighe, M., Lutfiyya, H. and Bauer, M. (2012), “An analysis of first fit heuristics for the virtual machine relocation problem”.
- Khani, H., Latifi, A., Yazdani, N. and Mohammadi, S. (2015), “Distributed consolidation of virtual machines for power efficiency in heterogeneous cloud data centers”, *Computers & Electrical Engineering*, Vol. 47, pp. 173–185.
- Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983), “Optimization by simulated annealing”, *Science (New York, N.Y.)*, Vol. 220 No. 4598, pp. 671–680.
- Koomey, J. (2011), “Growth in data center electricity use 2005 to 2010”, A report by *Analytical Press*, completed at the request of The New York Times, Vol. 9.
- Kumar, J. and Singh, A. K. (2018), “Workload prediction in cloud using artificial neural network and adaptive differential evolution”, *Future Generation Computer Systems*, Vol. 81, pp. 41–52.
- Kumar, M. R. V. and Raghunathan, S. (2016), “Heterogeneity and thermal aware adaptive heuristics for energy efficient consolidation of virtual machines in infrastructure clouds”, *Journal of Computer and System Sciences*, Vol. 82 No. 2, pp. 191–212.
- Kushida, K. E., Murray, J. and Zysman, J. (2011), “Diffusing the Cloud: Cloud Computing and Implications for Public Policy”, *Journal of Industry, Competition and Trade*, Vol. 11 No. 3, pp. 209–237.
- Lopez-Pires, F. and Baran, B. (2015), “Virtual machine placement literature review”, *arXiv preprint arXiv:1506.01509*.
- Luo, L., Wu, W., Tsai, W. T., Di, D. and Zhang, F. (2013), “Simulation of power consumption of cloud data centers”, *Simulation Modelling Practice and Theory*, Vol. 39, pp. 152–171.
- Malekloo, M. -H., Kara, N. and El Barachi, M. (2018), “An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments”, *Sustainable Computing: Informatics and Systems*, Vol. 17, pp. 9–24.
- Marotta, A., Avallone, S. and Kassler, A. (2018), “A Joint Power Efficient Server and Network Consolidation approach for virtualized data centers”, *Computer Networks*, Vol. 130, pp. 65–80.
- Nahhas, A., Bosse, S. and Turowski, K. (2018), “Load distribution strategies for a sustainable IT resources management”, in Drews, P., Funk, B., Niemeyer, P. and Xie, L. (Eds.), *Multikonferenz Wirtschaftsinformatik 2018, 2018-March*, Leuphana Universität Lüneburg Institut für Wirtschaftsinformatik, Lüneburg.
- Pinedo, M. (2012), *Scheduling: Theory, algorithms, and systems*, 4th ed., Springer, New York.
- Skiena, S. S. (1998), *The algorithm design manual: Text, Springer Science & Business Media*.
- Suresh, S. and Sakthivel, S. (2017), “A novel performance constrained power management framework for cloud computing using an adaptive node scaling approach”, *Computers & Electrical Engineering*, Vol. 60, pp. 30–44.
- Tchana, A., Son Tran, G., Broto, L., DePalma, N. and Hagimont, D. (2013), “Two levels autonomic resource management in virtualized IaaS”, *Future Generation Computer Systems*, Vol. 29 No. 6, pp. 1319–1332.
- Tesfatsion, S. K., Wadbro, E. and Tordsson, J. (2014), “A combined frequency scaling and application elasticity approach for energy-efficient cloud computing”, *Sustainable Computing: Informatics and Systems*, Vol. 4 No. 4, pp. 205–214.
- Vafamehr, A. and Khodayar, M. E. (2018), “Energy-aware cloud computing”, *The Electricity Journal*, Vol. 31 No. 2, pp. 40–49.
- Vitali, M., Pernici, B. and O’Reilly, U.-M. (2015), “Learning a goal-oriented model for energy efficient adaptive applications in data centers”, *Information Sciences*, Vol. 319, pp. 152–170.
- Wang, X., Du, Z., Chen, Y. and Li, S. (2008), “Virtualization-based autonomic resource management for multi-tier Web applications in shared data center”, *Journal of Systems and Software*, Vol. 81 No. 9, pp. 1591–1608.
- Xu, C. -Z., Rao, J. and Bu, X. (2012), “URL. A unified reinforcement learning approach for autonomic cloud management”, *Journal of Parallel and Distributed Computing*, Vol. 72 No. 2, pp. 95–105.
- Yoon, M. S., Kamal, A. E. and Zhu, Z. (2017), “Adaptive data center activation with user request prediction”, *Computer Networks*, Vol. 122, pp. 191–204.
- Zhang, Q., Metri, G., Raghavan, S. and Shi, W. (2014), “RESCUE: An energy-aware scheduler for cloud environments”, *Sustainable Computing: Informatics and Systems*, Vol. 4 No. 4, pp. 215–224.
- Zheng, X. and Cai, Y. (2011), “Energy-aware load dispatching in geographically located Internet data centers”, *Sustainable Computing: Informatics and Systems*, Vol. 1 No. 4, pp. 275–285.
- Zhou, H., Li, Q., Choo, K. -K. R. and Zhu, H. (2018), “DADTA: A novel adaptive strategy for energy and performance efficient virtual machine consolidation”, *Journal of Parallel and Distributed Computing*, Vol. 121, pp. 15–26.