A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments

Conference Paper · December 2015

DOI: 10.1109/ICMLA.2015.205

A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments

Mehmet Demirci
Department of Computer Engineering
Gazi University
Ankara, Turkey
Email: mdemirci@gazi.edu.tr

***Abstract*—Ensuring energy efficiency in data centers is a crucial objective in modern cloud computing because it reduces operating costs and complies with the goals of green computing. Researchers strive to develop optimal policies for resource management in the cloud, which has many components such as virtual machine placement, task scheduling, workload consolidation, and so on. Machine learning has a major role to play in these efforts. In this paper, we provide a detailed survey of recent works in the literature which have employed machine learning (ML) to offer solutions for energy efficiency in cloud computing environments. We also present a comparative classification of the proposed methods. Furthermore, we enrich this survey by studying non-ML proposals to energy conservation in data centers, and also how ML has been applied towards other objectives in the cloud.**

***Index Terms*—Energy Efficiency, Resource Management, Data Centers, Cloud Computing, Machine Learning**

I. Introduction

Cloud computing has matured and established itself as an indispensable part of information technology in recent years. As a computational paradigm enabling economies of scale, when configured and utilized effectively, cloud computing offers substantial benefits in terms of computation power while reducing costs and saving energy. Large data centers are the places where the idea of cloud computing comes to life. Through virtualization technology, it becomes possible for many users to share data center resources and services, thus avoiding having to set up their own infrastructure to do things which can be done on the cloud.

Resource management in the cloud (or in data centers) signifies the decisions governing the allocation of processing power (or CPU time), memory, storage, network bandwidth etc. among different services requesting them. Successful resource management may mean different things for cloud providers depending on their priorities and objectives, which may involve balancing the load, maximizing revenue, minimizing response time or the usage of server resources, bandwidth, electricity etc.

According to previous research [1], approximately 15% of data center costs are electrical utility costs. Reducing these costs may be possible through innovations in hardware technology, as well as efficient operation and management of data center networks. Intelligent resource management with the objective of maximizing energy efficiency can help to increase savings. Therefore, developing optimal or near-optimal strategies for minimizing energy usage is a worthwhile endeavor.

Energy efficiency in cloud computing has been a popular research topic over the last decade. A number of works, some of which will be discussed in Section II, have proposed different kinds of optimization solutions to the problem of minimizing energy costs in cloud computing environments. There are also many examples of applying machine learning techniques to resource provision and management in the cloud with various objectives. This survey paper focuses on the intersection of the two aforementioned groups of works: We provide a survey of machine learning-based proposals to reducing energy usage in data centers. Our goal is to shed light on this branch of energy efficiency research, present the state of the art to machine learning researchers, and assist them in developing novel approaches that can produce successful solutions.

The rest of the paper is organized as follows. In Section II, we take a look at cloud resource management from two directions: optimizations for energy-efficient resource management, and machine learning solutions targeting objectives besides energy efficiency. In Section III, we converge on the main topic in this work: how machine learning has been used by researchers to improve energy efficiency in the cloud. Section IV includes a comparative discussion of the research reviewed in this paper. Finally, Section V summarizes the paper and gives possible directions for future work in this area.

II. Energy Efficiency and Machine Learning in the Cloud

This section provides context for the main topic of the paper by first introducing other methods besides machine learning for energy efficiency in the cloud, and then reviewing some machine learning applications to general cloud resource man-

This work was supported in part by TUBITAK Grant 114C093.

978-1-5090-0287-0/15 $31.00 © 2015 IEEE
DOI 10.1109/ICMLA.2015.205

agement in the literature, where the objective is not directly energy-related.

A. Optimizing for Energy Efficiency

Technologies such as parallel programming and virtualization enabled multiple users and applications sharing computing resources, i.e., multi-tenancy. Server clusters, although not at the scale of modern data centers, are environments where multi-tenancy should be supported. Heath et al. [2] focused on the problem of distributing client requests to the servers in a heterogeneous cluster so that a good tradeoff between throughput and energy savings will be found. They designed a cluster with the ability to configure itself to optimize for a variety of metrics (or a combination thereof), among which energy consumption was featured and reducing it was chosen as a main objective. They developed analytical models for request distributions and resource utilization, and then used simulated annealing to find a request distribution that minimizes power-to-throughput ratio. Their method offered over 40% savings in energy consumption.

Chen et al. [3] designed algorithms to minimize the number of running servers via dynamic provisioning and distribute the load to these servers in such a way that saves up to 30% energy. Urgaonkar et al. [4] devised an online control algorithm to perform admission control and resource management in a data center using Lyapunov Optimization. They formulated stochastic optimization problems whose solutions maximize a joint utility combining application throughput and the amount of energy saved.

Beloglazov et al. [5], [6] proposed policies for energy-efficient cloud resource allocation and scheduling algorithms to meet quality of service demands while maintaining low power usage. They made use of dynamic virtual machine (VM) allocation and live migration (movement from one physical node to another) to achieve reduced energy consumption. The algorithm for VM placement is a Modified Best Fit Decreasing algorithm, and the algorithm for deciding which VMs to move looks to minimize the number of VM migrations needed to lower CPU utilization below a threshold at all hosts.

Gandhi et al. [7] employed a combination of predictive and reactive resource provisioning to satisfy service-level agreements (SLA) while saving up to 35% energy. The proposed system analyzes long-term demand histories to establish a base workload, feeds this base workload to a predictive controller that allocates just enough resources to meet SLA requirements, and augments it with a simple reactive controller to handle cases where actual demand is higher than the predicted demand.

Van et al. [8] developed dynamic VM provisioning and placement managers that strive to ensure SLA compliance while reducing energy consumption, and place VMs on the minimum number of physical machines via live migration. Consolidating tasks in a data center can reduce the number of servers needed, yet released resources may still consume energy in the idle state. Lee and Zomaya [9] considered idle power draw in addition to active energy consumption and developed heuristics to minimize total energy consumption.

Xu and Fortes [10] used a genetic algorithm supported with fuzzy multi-objective optimization to simultaneously minimize the amount of wasted resources, energy consumption and thermal dissipation costs.

Shen et al. [11] presented CloudScale, a prediction-based online system for adaptive cloud resource allocation. Their resource demand predictor uses a fast Fourier transform to identify a signature that can be used to estimate future demands. If a signature is not found, the predictor then employs a discrete-time Markov chain. Furthermore, CloudScale uses preemptive VM migration to avoid conflicts, along with dynamic voltage and frequency scaling to save energy. The result is that CloudScale can save 8-10% total energy consumption.

Heller et al. [12] introduced ElasticTree, a power manager with a focus on the data center network elements (links and switches). ElasticTree monitors traffic conditions in the data center, and simply turns off the switches and links if they are not needed. The authors used a combination of linear programming, greedy bin-packing, and topology-aware heuristic approaches to achieve up to 50% reduction in network energy usage.

Berl et al. [13] provided a more comprehensive survey of energy-efficient cloud computing solutions.

B. Machine Learning for the Cloud

There exist many examples of machine learning based solutions to various resource management problems in the cloud. In this subsection, we will discuss several such solutions with varying objectives. It is important to note that although these solutions were not explicitly designed for energy efficiency, some of them may reduce energy consumption as a side effect.

Liao et al. [14] used machine learning to find the best configuration for memory prefetchers. They employed a variety of algorithms including nearest neighbor, naive Bayes, C4.5 decision tree, Ripper, support vector machines, logistic regression, multi-layer perceptron, and radial basis function.

Bodik et al. [15] applied statistical machine learning models, such as linear and LOESS regression, to optimal control for data centers.

Predicting how much time and resources will be spent by applications is necessary to be able to schedule jobs efficiently. Matsunaga and Fortes [16] viewed this as a supervised machine learning problem and extended the Predicting Query Runtime algorithm [17] to the regression problem, calling their modified method PQR2. Using Weka, they compared PQR2 to a group of machine learning algorithms including k-nearest neighbors, linear regression, decision tree, radial basis function, and support vector machine. They showed that PQR2 offers the best accuracy among these algorithms. In another work on web applications, Jiang et al. [18] combined machine learning (linear regression) with time series analysis to predict the number of requests and decide whether to increase or decrease the number of active VMs.

Islam et al. [19] utilized error correction neural network and linear regression along with sliding window to predict resource usage patterns in the cloud. They evaluated prediction accuracy using data generated by running TPC-W [20] on Amazon EC2 and demonstrated the effectiveness of their approach. They also tried different sliding window sizes and concluded that neural network with optimal sliding window size performs better than linear regression. Gong et al. [21] developed another system called PRESS that predicts cloud resource demands to perform elastic resource scaling using statistical machine learning methods. Bankole and Ajila [22] evaluated their resource demand prediction models built using support vector machine, neural network and linear regression. They concluded that SVM attains the best results in terms of response time and throughput.

Maximizing profits is a crucial goal for cloud providers. To this end, Xiong et al. [23] proposed SmartSLA, a resource management system that consolidates multiple VMs into a single physical machine to reduce costs while complying with tenant SLAs. Machine learning comes into play in learning a model describing how different resource allocations to clients correspond to potential profits. After this, the module responsible for resource allocation uses the learned model to adjust allocations and maximize profits. The authors realized that simple linear regression was unsatisfactory for their objectives, so they turned to the regression tree model and added a boosting approach called additive regression to decrease the prediction error.

Several researchers applied reinforcement learning to resource allocation and management in the cloud. Xu et al. [24] took a unified reinforcement learning approach to determine the optimal configurations for VMs in a cloud computing environment. Dutreilh et al. [25] integrated reinforcement learning solutions in an automated controller for the cloud. Their workflow presented three key components for tuning the model: Q-function initialization, convergence speedups, and performance model change detection.

Barrett et al. [26] proposed a parallel Q-learning approach to reduce the amount of time required to zero in on the optimal resource scaling policy during online learning. The authors stated that their approach was the first application of parallelized reinforcement learning to improving convergence times. Their evaluation which used multiple learning agents (varying from 2 to 10) showed that parallelization was able to provide meaningful reduction in convergence times. Rao et al. [27] viewed cloud resource management as a task suitable for distributed learning, and utilized reinforcement learning to develop a mechanism where each VM acts as an autonomous agent in the learning process.

Kundu et al. [28] applied refinements to artificial neural network and support vector machine algorithms to model the relationship between application performance and the resources allocated to the VM hosting the application. They argue that their results are significantly better than those provided by non-refined machine learning methods or regression approaches. Huang et al. [29] employ support vector regression technique in conjunction with a genetic algorithm to reduce application service response times in the cloud.

Virtual network (VN) embedding [30] is closely related to cloud resource management. In data centers, the substrate network in the data center network, and incoming VN requests must be met with efficient resource allocations. Mijumbi et al. [31] design a reinforcement learning algorithm to perform substrate resource management. Their main objective is increasing VN acceptance ratio, i.e, the fraction of incoming VN requests that are successfully answered.

III. Machine Learning Solutions to Energy-Efficient Cloud Resource Management

In this section, we review a number of machine learning-based proposals for enabling energy efficiency in cloud computing environments.

There are many factors influencing energy consumption in data centers such as power distribution, the heat produced by data center operations and resulting cooling costs, and the management of computing load [32]. Most of the current solutions offered for energy efficiency in data centers focus on optimally distributing the computing load so that the minimum number of machines will be activated to satisfy application demands.

Vasic et al. [33] proposed DejaVu, a resource management system for the cloud that learns from the results of previous resource allocations. The learning phase of DejaVu takes about a week of service use when workloads and their corresponding resource allocations are identified. In actual use, DejaVu automatically classifies each a workload to check if it matches a previously encountered workload. Depending on the classification result, it either reuses the previous allocation, or orders the service to reconfigure itself. The authors argued that their efficient mechanism for adapting to new workloads would result in lowered energy costs as it allows load consolidation.

Demand forecasting is a key problem in data center management, and good forecasting techniques can lead to optimal allocation strategies that minimize energy consumption. Prevost et al. [34] utilized neural network and auto-regressive linear prediction to forecast future demand profiles. Performance results of a multi-layer perceptron model and a linear auto-regressive predictor were compared. The authors concluded that the linear predictor was able to produce a more accurate model. Similarly, Duy et al. [35] turn to a neural network based predictor for load forecasting in the cloud. They used the results of the forecast to turn off unused servers and conserve energy. Using historical demand data to train their system, the authors were able to demonstrate that their scheduling algorithm was able to save over 40% energy.

Dabbagh et al. [36] developed another framework for predicting future virtual machine requests and associated resource requirements. This framework uses this predictor to reduce energy consumption by putting unneeded machines into sleep mode. The techniques used are k-means for clustering, and stochastic Wiener filter for workload prediction. Using Google

TABLE I
CATEGORIZATION OF ML PROPOSALS TO ENERGY EFFICIENCY IN THE CLOUD

Year	Authors	Learning Model	Objective (Energy saving method)
2007	Tesauro et al. [40]	Hybrid (Reinforcement + Supervised)	Adjusting CPU frequency
2007	Tesauro et al. [41]	Hybrid (Reinforcement + Supervised)	Power-aware server allocation
2010	Duy et al. [35]	Supervised	Intelligent scheduling to turn off unused servers
2010-11	Berral et al. [38], [39]	Supervised	Power-aware task scheduling and consolidation
2011	Prevost et al. [34]	Supervised	Load prediction leading to optimal resource allocation
2011	Chen et al. [32]	Reinforcement	Spatially-aware load placement to reduce cooling costs
2012	Vasic et al. [33]	Supervised	Workload classification leading to load consolidation
2014	Dabbagh et al. [36]	Unsupervised	Request forecasting to put unused machines to sleep

traces collected over 29 days, the authors showed that their framework achieved near-optimal energy efficiency.

A few solutions go beyond this simple objective of activating the minimum number of physical machines. Chen et al. [32] considered power distribution and cooling in deciding where to place the computing load. They used a model-based reinforcement learning approach to learn the thermal distribution resulting from different workload placements, and then predict the thermal distribution of incoming workloads under various placement alternatives in order to pick the optimal one. The authors called their method spatially-aware workload management (SpAWM), and deployed it using VMWare's ESXServer virtualization infrastructure. They used the Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library (PyBrain) [37] to implement their online algorithm, which combines neural networks and reinforcement learning Evaluation results showed a $2 - 3°$C decrease in temperature, which led to saving $13\% - 18\%$ cooling energy.

Machine learning methods can be used as complementary tools in building holistic solutions to energy efficiency in the cloud. Berral et al. [38], [39] employed supervised machine learning methods to predict resource consumption by different tasks and SLA-related metrics such as response time for a given workload, and integrated their predictor in a system that performs power-aware task scheduling and consolidation. The algorithms they used were linear regression to predict CPU usage at each host, and a more complex machine learning algorithm called M5P to predict power consumption. Experiments were carried out using real workloads, and demonstrated that their methods offered substantial power savings while slightly decreasing performance.

Certain solutions in the literature apply machine learning at the level of individual servers in the data center. Tesauro et al. [40] presented a reinforcement learning approach to simultaneously manage performance and power consumption by intelligently adjusting CPU frequency online. Their approach is tunable in that it uses a simple objective function that subtracts power consumption multiplied by a modifiable coefficient from a performance-based utility. They employed the Sarsa(0) update rule, and trained a multilayer perceptron neural network. Their paper detailed some specific innovations in applying reinforcement learning and justified their use through experimental evaluation.

In another work, Tesauro et al. [41] proposed a hybrid reinforcement learning approach supported with neural networks, this time for enabling intelligent server allocation decisions. By combining the flexibility of reinforcement learning, which does not require explicit models, and the ability of model-based policies to quickly reach high performance, they were able to attain successful results while avoiding the potential performance problems associated with online reinforcement learning.

IV. DISCUSSION

Table I presents a categorization of the proposals summarized in the previous section. We observe that supervised methods are widely used, and among these artificial neural network and linear regression models are the most common. A couple of solutions [40], [41] augment reinforcement learning with supervised techniques (such as multi-layer perceptron) to try and get the best of both worlds.

Most of the proposals have employed machine learning in predicting future workloads in the data center. Once this prediction is made, a variety of algorithms can be used to maximize energy conservation. A straightforward goal is to minimize the number of servers needed and simply turn off unused machines or put them in sleep mode. A more creative approach to resource management involves factoring thermal distribution into workload placement decisions. Moving the computing load in such a way that the resulting thermal distribution will necessitate the minimum amount of cooling energy was proved to offer meaningful power conservation [32].

An interesting area of research that is lacking in the literature is the interaction between seemingly conflicting cloud resource management objectives, such as energy efficiency and fault tolerance. Energy conservation is usually enabled through server consolidation where computing load is placed on the lowest number of servers possible. However, in case of a failure, this policy could lead to weakened resilience and fault tolerance. Although redundancy is typically not a friend of energy efficiency, selective server and virtual machine replication has the potential to preserve energy efficiency while maintaining resilient operation in the data center. Machine learning could well be the key in finding intelligent ways to determine the scope of such replication, that is, selecting where and when to replicate.

V. Concluding Remarks

In this paper, we have provided an extensive review of machine learning applications towards energy efficient management of cloud data centers. The most common use of machine learning in this context is for predicting future resource demands. Accurate estimation of these demands can allow cloud providers to develop intelligent resource management policies which rely on task scheduling and consolidation to turn on the minimum number of machines in the data center, thus conserving energy. In addition, researchers have found ways to reduce cooling costs by paying attention to the thermal distribution resulting from different workload placements. On the scale of individual machines, machine learning has been used to save energy by optimally configuring CPU frequency.

Many works in the literature combine energy efficiency with a performance objective such as SLA satisfaction, or a monetary objective such as maximized profits, and seek a joint optimization or study the tradeoffs. One suggestion for potential future work is pairing energy efficiency with a different sort of objective such as reliability, survivability, fault tolerance, or security. Another avenue for future work in this area could be energy-efficient configuration of both virtual network and data center network topologies using machine learning techniques.

References

[1] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68–73, 2008.

[2] T. Heath, B. Diniz, E. V. Carrera, W. Meira Jr, and R. Bianchini, "Energy conservation in heterogeneous server clusters," in *Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming*. ACM, 2005, pp. 186–195.

[3] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services." in *NSDI*, vol. 8, 2008, pp. 337–350.

[4] R. Urgaonkar, U. C. Kozat, K. Igarashi, and M. J. Neely, "Dynamic resource allocation and power management in virtualized data centers," in *Network Operations and Management Symposium (NOMS), 2010 IEEE*. IEEE, 2010, pp. 479–486.

[5] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2010, pp. 826–831.

[6] A. Beloglazov, J. Abawajy, and R. Buyya, "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, 2012.

[7] A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah, "Minimizing data center sla violations and power consumption via hybrid resource provisioning," in *Green Computing Conference and Workshops (IGCC), 2011 International*. IEEE, 2011, pp. 1–8.

[8] H. N. Van, F. D. Tran, and J.-M. Menaud, "Performance and power management for cloud infrastructures," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. IEEE, 2010, pp. 329–336.

[9] Y. C. Lee and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," *The Journal of Supercomputing*, vol. 60, no. 2, pp. 268–280, 2012.

[10] J. Xu and J. A. Fortes, "Multi-objective virtual machine placement in virtualized data center environments," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*. IEEE, 2010, pp. 179–188.

[11] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, 2011, p. 5.

[12] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "Elastictree: Saving energy in data center networks." in *NSDI*, vol. 10, 2010, pp. 249–264.

[13] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Q. Dang, and K. Pentikousis, "Energy-efficient cloud computing," *The computer journal*, vol. 53, no. 7, pp. 1045–1051, 2010.

[14] S.-w. Liao, T.-H. Hung, D. Nguyen, C. Chou, C. Tu, and H. Zhou, "Machine learning-based prefetch optimization for data center applications," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. ACM, 2009, p. 56.

[15] P. Bodık, R. Griffith, C. Sutton, A. Fox, M. Jordan, and D. Patterson, "Statistical machine learning makes automatic control practical for internet datacenters," in *Proceedings of the 2009 conference on Hot topics in cloud computing*, 2009, pp. 12–12.

[16] A. Matsunaga and J. A. Fortes, "On the use of machine learning to predict the time and resources consumed by applications," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2010, pp. 495–504.

[17] C. Gupta, A. Mehta, and U. Dayal, "Pqr: Predicting query execution times for autonomous workload management," in *Autonomic Computing, 2008. ICAC'08. International Conference on*. IEEE, 2008, pp. 13–22.

[18] J. Jiang, J. Lu, G. Zhang, and G. Long, "Optimal cloud resource autoscaling for web applications," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*. IEEE, 2013, pp. 58–65.

[19] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 155–162, 2012.

[20] "TPC-W," http://www.tpc.org/tpcw/, [Online; accessed 26-August-2015].

[21] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*. IEEE, 2010, pp. 9–16.

[22] A. A. Bankole and S. A. Ajila, "Predicting cloud resource provisioning using machine learning techniques," in *Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on*. IEEE, 2013, pp. 1–4.

[23] P. Xiong, Y. Chi, S. Zhu, H. J. Moon, C. Pu, and H. Hacigümüş, "Intelligent management of virtualized resources for database systems in cloud environment," in *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 2011, pp. 87–98.

[24] C.-Z. Xu, J. Rao, and X. Bu, "Url: A unified reinforcement learning approach for autonomic cloud management," *Journal of Parallel and Distributed Computing*, vol. 72, no. 2, pp. 95–105, 2012.

[25] X. Dutreilh, S. Kirgizov, O. Melekhova, J. Malenfant, N. Rivierre, and I. Truck, "Using reinforcement learning for autonomic resource allocation in clouds: Towards a fully automated workflow," in *ICAS 2011, The Seventh International Conference on Autonomic and Autonomous Systems*, 2011, pp. 67–74.

[26] E. Barrett, E. Howley, and J. Duggan, "Applying reinforcement learning towards automating resource allocation and application scalability in the cloud," *Concurrency and Computation: Practice and Experience*, vol. 25, no. 12, pp. 1656–1674, 2013.

[27] J. Rao, X. Bu, C.-Z. Xu, and K. Wang, "A distributed self-learning approach for elastic provisioning of virtualized cloud resources," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on*. IEEE, 2011, pp. 45–54.

[28] S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta, "Modeling virtualized applications using machine learning techniques," in *ACM SIGPLAN Notices*, vol. 47, no. 7. ACM, 2012, pp. 3–14.

[29] C.-J. Huang, Y.-W. Wang, C.-T. Guan, H.-M. Chen, and J.-J. Jian, "Applications of machine learning to resource management in cloud computing," *International Journal of Modeling and Optimization*, vol. 3, no. 2, p. 148, 2013.

[30] A. Fischer, J. F. Botero, M. Till Beck, H. De Meer, and X. Hesselbach, "Virtual network embedding: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 4, pp. 1888–1906, 2013.

[31] R. Mijumbi, J.-L. Gorricho, J. Serrat, M. Claeys, F. De Turck, and S. Latré, "Design and evaluation of learning algorithms for dynamic resource management in virtual networks," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*. IEEE, 2014, pp. 1–9.

[32] H. Chen, M. Kesavan, K. Schwan, A. Gavrilovska, P. Kumar, and Y. Joshi, "Spatially-aware optimization of energy consumption in con-

solidated data center systems," in *ASME 2011 Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems*. American Society of Mechanical Engineers, 2011, pp. 461–470.

[33] N. Vasić, D. Novaković, S. Miučin, D. Kostić, and R. Bianchini, "Dejavu: accelerating resource allocation in virtualized environments," in *ACM SIGARCH Computer Architecture News*, vol. 40, no. 1. ACM, 2012, pp. 423–436.

[34] J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, "Prediction of cloud data center networks loads using stochastic and neural models," in *System of Systems Engineering (SoSE), 2011 6th International Conference on*. IEEE, 2011, pp. 276–281.

[35] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Performance evaluation of a green scheduling algorithm for energy savings in cloud computing," in *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1–8.

[36] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, "Energy-efficient cloud resource management," in *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. IEEE, 2014, pp. 386–391.

[37] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber, "Pybrain," *The Journal of Machine Learning Research*, vol. 11, pp. 743–746, 2010.

[38] J. L. Berral, Í. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres, "Towards energy-aware scheduling in data centers using machine learning," in *Proceedings of the 1st International Conference on energy-Efficient Computing and Networking*. ACM, 2010, pp. 215–224.

[39] J. L. Berral, R. Gavalda, and J. Torres, "Adaptive scheduling on power-aware managed data-centers using machine learning," in *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*. IEEE Computer Society, 2011, pp. 66–73.

[40] G. Tesauro, R. Das, H. Chan, J. Kephart, D. Levine, F. Rawson, and C. Lefurgy, "Managing power consumption and performance of computing systems using reinforcement learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 1497–1504.

[41] G. Tesauro, N. K. Jong, R. Das, and M. N. Bennani, "On the use of hybrid reinforcement learning for autonomic resource allocation," *Cluster Computing*, vol. 10, no. 3, pp. 287–299, 2007.