As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 9(03): 123-131, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0903018
Available online at: http://www.ijmtst.com/vol9issue03.html

Virtualized Resource Management in Cloud Data Center: A Systematic Survey

Madhukar Shelar[1] | Archana Bachhav[2]

[1]Department of Computer Science, KRT Arts, BH Commerce and AM Science (KTHM) College, Nashik, India
[2]Department of Computer Science, KSKW Arts, Science and Commerce College, Nashik, India

To Cite this Article
Madhukar Shelar and Archana Bachhav. Virtualized Resource Management in Cloud Data Center: A Systematic Survey. International Journal for Modern Trends in Science and Technology 2023, 9(03), pp. 123-131. https://doi.org/10.46501/IJMTST0903018.

Article Info
Received: 16 February 2023; Accepted: 11 March 2023; Published: 13 March 2023.

ABSTRACT

The demand for computing resources like processing speed, memory, storage, and bandwidth is continually rising in cloud data centres. Cloud service providers have put up data centres across the world to meet the resource requirement of web applications. Thousands of extremely powerful servers are often installed in these data centres. The server virtualization approach is used to allocate resources during the deployment and provisioning of applications. In the world of cloud computing, server virtualization has emerged as a crucial component. It makes it possible to run multiple Virtual Machines (VMs) concurrently on top of a single Physical Machine (PM) or Server. Researchers have developed different VM placement and consolidation algorithms with trade-off between various conflicting performance parameters such as power consumption, SLA (Service Level Agreement) violation, application response time, number of VM migrations etc. Several researchers have researched this issue and offered a number of strategies to improve the values of one or more competing performance parameters. This research paper presents taxonomy and extensive systematic survey of several research papers on virtualized resource management in cloud environment to meet these performance parameters.

KEYWORDS: Cloud Data Center, Cloud Service Provider, Virtual Machine, Virtualization, Virtualized Resource Management, SLA Violation.

1. INTRODUCTION

The more recent trend in distributed computing that makes it easier to supply computing resources, platforms, and applications as a service is called cloud computing. Cloud computing is a very well-liked computing model that allows consumers to access virtualized, scalable resources as services in recent years. Internet users today employ a wide range of cloud-based goods and applications, including Amazon EC2, Google App Engine, Microsoft Azure, IBM Blue Mix and many others. Numerous consumers and corporations also use the internet to deploy their web applications on cloud data centres. In order to satisfy the growing demands of customers for computing resources such as processing power, memory, storage etc. large cloud data centres have been established across the globe containing thousands of servers by cloud service providers. To address the issue of poor resource utilization, multiple applications could be hosted on single Physical Machine(PM) by creating several Virtual Machines (VM) using the technique called server virtualization.

Table 1. Taxonomy of resource management approaches

Cost	• *Saving energy cost* • *Reduction of operational cost* • *Reduction of resource cost*
Performance	• *Load balancing* • *Avoiding resource contention* • *Minimization of VM migrations*
Availability	• *Multiple instances of applications in a VM* • *Multiple copies of VMs*
Traffic	• *Placement of related apps at close proximity* • *Minimization of network utilization*
Energy	• *Maximizing server utilization* • *Lower threshold to keep servers at idle state.* • *Idle servers at power saving state*
SLA Violation	• *Upper thresholds at servers* • *Live VM migration*
Future Workload	• *Prediction on past workload* • *Prediction on statistical analysis*
Fragment	• *Allocation of resources as per peak load* • *Reducing skewness of resource allocation*
Scaling	• *Vertical scaling* • *Hybrid scaling*

Researchers have proposed different techniques for management of computing resources using various approaches for resource management that are based on cost, performance, availability, traffic, energy, SLA violation, future workload, fragmentation and scalability. Table 1 shows the taxonomy of cloud resource management approaches in which various techniques or solutions to achieve appropriate approaches are defined. Research work based on these various resource management approaches was reviewed and presented in next section.

STRUCTURE OF PAPER

The paper is organized as follows: Section I puts the introduction of the paper along with various techniques to achieve suitable cloud resource management approaches. In Section II we present the extensive survey of research work on virtualized resource management in cloud data centers by considering several methodoliges. Section III remarks the conclusion and future scope followed by references.

2. RESOURCE MANAGEMENT APPROACHES

Cost based management

For some researchers, cost savings is the main consideration for more effective use of computing resources. By reducing the amount of resources used, virtualized cloud architecture significantly contributes to cost reduction [1]. Customers are also focused on minimising the amount of cloud resources consumed in order to lower their bill. As a result, both the service provider at the virtualization level and the customer at the application level need take into account cloud resource management [2].

In order to reduce total operational costs, researchers in [3], [4], [5] adopted the BFD (Best Fit Decreasing) algorithm, which was initially designed for the bin packing problem, for VM to PM mapping. Each VM is placed on a PM that results in the least increase in energy usage after the BFD algorithm sorts all VMs in descending order of their CPU utilisations. The overhead associated with this strategy is sorting VMs. In order to obtain cost savings for better usage of computer resources utilising any suitable resource management policy, Hyser et al. [6] offered a high level overview of VM placement and proposed a system architectural design of an autonomous VM placement. Nevertheless, this strategy does not take into account the VM consolidation element, which could result in performance degradation. Speitkamp et al. [7] suggested a VM consolidation strategy that combines data analysis to characterise variations in workload traces and apply algorithms to optimally assign VMs to target servers. In this study, the VM consolidation method is employed to make efficient use of computing resources and lower data centre running expenses. Costs for energy use, storage, upkeep, management, and server purchases are all included. This approach gives decisions that seek to reduce investment and operating costs extra support. Nevertheless, it has not addressed on-demand plans to offer adequate resources.

The optimal virtual machine placement algorithm put out by Chaisiri et al. [8] can reduce the cost associated with on-demand and reservation-based resource supply schemes. The method operates in two stages: the first determines the number of virtual machines (VMs) that are in demand, and the second determines the number of VMs that are allotted. This algorithm takes into account all potential pricing and demands. But, in the actual world, this method cannot scale as the number of potential demands and costs increases. Authors have created an autonomous resource manager in [9], [10] in which a specified set of VM classes are based on resource capacity. Each VM class has a defined set of CPU and memory specifications. Depending on the workload at hand, a VM must be selected for applications from a list of preset VM classes. It seeks to maximise a global utility function that incorporates the level of SLA adherence and the operating expenses. The biggest disadvantage of this strategy is the potential for excessive resource allocation for apps in virtual machines.

Performance based management

Another factor that many researchers take into account for the VM placement issue is performance. This can be addressed by distributing load of applications evenly among all physical servers which is termed as load balancing policy [11]. The research in [6] concentrated on developing a framework with a load balancing policy, in which data centre load is distributed among all available physical servers and resource usage is balanced to the greatest extent across all resource types, such as processing power, memory, bandwidth, storage, etc. Nevertheless, this effort does not employ any dynamic VM consolidation techniques, and as a result, peak load performance may suffer. In [12], the Eucalyptus performance-aware open source software framework is described. It implements Infrastructure as a Service (IaaS) and offers academic research organisations a modular platform for experimental instrumentation and investigation. Round robin VM scheduling, used by Eucalyptus, uniformly distributes the workload of virtual machines across all servers in the data centre. It improves the performance of running programmes, but in fact it is very wasteful because each time the scheduler distributes VMs to a processor and consequently the power consumption grows by its utmost potential [13].

The PACMan VM consolidation manager, which Roytman et al. [14] devised, reduces power usage while maintaining performance. By suggesting a system that chooses VMs with the least amount of interference from one another and groups them on the same server, it has concentrated on the performance loss that happens as a result of resource contention among VMs. PACMan arranges VM combinations so as to cause the least amount of performance damage. Although choosing the optimum group of VMs to group them with is an NP-Complete problem, PACMan uses a close to optimal yet approximation technique. Unfortunately, this method provides the best answer when a group of virtual machines are prepared for deployment, but it does not do so for placing virtual machines as they arrive for deployment in the cloud. Numerous current approaches to increasing application performance have a strong emphasis on resource usage, which may be unjust to some apps. Resource management techniques that take into account the equal satisfaction of all applications employing suitable resource allocation have been developed by Carrera et al. [15].

By running many instances of a VM on various servers and dividing incoming requests among them, it is possible to increase the speed of applications. The VM placement methodology was proposed by Goudarzi et al. [16] and is based on dynamic programming and local search techniques. Each VM is assigned a certain number of instances through dynamic programming, which places them on the relevant servers. By turning off idle servers, the local search technique reduces energy costs. This method helps the cloud provider make better use of the servers by lowering the resource requirements for each VM instance. However, keeping VM instances consistent may require extra bandwidth in the data centre.

Availability based management

The availability of cloud applications is another crucial component in cloud data centres, in addition to cost savings and performance. To boost availability in the event of failure, Goudarzi et al. [16] suggested a method for maintaining several copies of VM and placing them on various servers. Each copy requires fewer resources, which makes the server more effectively used. The challenge of preserving consistency and coordination between numerous copies of VMs, as well as failure recovery, is not the focus of this study. Applications that need higher availability can be distributed among several servers, however a cluster of computers that are near together physically could have failures. For instance, the failure of a rack or racks due to a network failure could bring down all servers connected to that rack or those racks, which could cause a data centre to fail. Shouraboura et al.[17] established the Virtual Cloud Model data structure by altering voronoi diagrams that distribute applications over remote servers to boost availability at the expense of latency in order to assure the continued operation of such applications. The deployment of individual VMs in which one or more applications are deployed is the focus of current approaches. Yet, many applications that are deployed in the cloud have many virtual machines (VMs) clustered together for availability or due to their multiform, multitier nature [18], [19]. A novel scheduling technique was proposed by Jammal et al. [20] that takes into account the interdependencies and redundancies between application components, their failure scopes, their communication delay tolerance, and resource utilisation requirements. They also presented the impact of application placement strategy on high availability of services to end users. The method looks at Mean Time To Repair (MTTR) and recovery in addition to Mean Time To Failure (MTTF) to measure the downtime.

Traffic based management

Public clouds are expanding quickly nowadays, with millions of servers spread across numerous data centres in numerous nations and continents. In order to facilitate cross-application communication, such public clouds must maintain a network with ample bandwidth. For groups of applications that frequently communicate, cloud providers must maintain the shortest possible communication paths. As a result, VM placement can be optimised by taking their traffic patterns into account. Meng et al. [21] introduce the traffic aware virtual machine placement to increase network scalability. In this study, physical hosts or host machines near to each other are allocated virtual machines with high mutual bandwidth utilisation. It would be challenging to estimate the cost of communication or traffic between two virtual computers because the traffic load across virtual machines is essentially dynamic in nature. Biran et al. [22] suggested VM placement algorithms that take into account the point of traffic determination between two VMs and attempt to satisfy the expected communication requirements while remaining unaffected by dynamic traffic changes.

Researchers have suggested a dynamic VM placement algorithm in [23], [24] to reallocate VMs among servers in a data centre based on the data center's current traffic matrix. Virtual machine migrations may result in dynamic changes to a data center's traffic matrix. A traffic matrix is built using the traffic that is sent back and forth between each pair of VMs. The ideal TVMPP (Traffic-aware VM Placement Problem) solution, as defined by Meng et al. [21], generates VM-to-PM mappings with the aid of the VM traffic matrix and the physical server communication cost matrix, leading to the lowest possible aggregate traffic rates at each network switch. The number of hops on the network's routing path determines the communication cost between any two VMs.

Enrgy based management

Data centres are strong infrastructure and communication technology facilities that constantly change in terms of size, complexity, and power consumption, while also changing in terms of the complexity of user requirements [25]. Cloud service providers make it a priority to meet all of their clients' needs while keeping them separate from the underlying infrastructure. In data centre deployments, high performance has been the main priority, and this need has been met without much consideration for energy consumption [26]. According to J. Kaplan et al. [27] technical assessment, a typical data centre uses as much energy as 25,000 residences. As a result, resource management must switch from performance-based optimization to energy-aware optimization while maintaining performance. Energy conservation has a critical role in reducing CO2 and greenhouse gas (GHG) emissions as well as power costs [28], [29]. The Dynamic Voltage and Frequency Scaling (DVFS) idea, which enables servers to run at different voltage and frequency combinations to reduce processor energy consumption, can be used to reduce the overall energy consumption of servers [30], [31]. With the least amount of performance loss, energy consumption is reduced. Feng et al. [32] used DVFS techniques to reduce the frequency by 400 MHz while only suffering a 5% performance hit. With the use of live VM migration and the development of power-aware scheduling approaches, researchers have improved system efficiency at the expense of a negligible performance overhead. Wu et al.[33] devised the DVFS-based scheduling system for cloud data centres maximises server usage and, as a result, uses little energy. An energy-efficient method was provided by Calheiros et al. [34] that uses a DVFS module and an intelligent scheduler for CPU-intensive workloads.

Reducing the number of active (running) physical servers in a data centre is one way to reduce energy usage [8], [35], [36]. The processing power of the CPU, followed by the memory, accounts for the majority of the power needed by a server, according to data published by Intel Labs [37]. As a result, the data centre hosts' power consumption increases linearly as CPU utilisation rises. Beloglazov et al. [3] proposal called for data centre resources to be allocated in an energy-conscious manner based on CPU consumption while adhering to strong consolidation standards. Several energy-conscious research studies solely take into account the processing power of the CPU when calculating the energy consumption of servers. Due to the nature of the workload, other resources like RAM and bandwidth could, however, constitute a bottleneck. By introducing an ecoCloud, a self-organizing and adaptive strategy for the VM consolidation with the objectives of energy saving and honouring SLA, Mastroianni et al. [38] extended the VM consolidation challenge on RAM in addition to CPU.

The majority of these study made the assumption that all physical servers in data centres are uniform, although server configurations in data centres vary widely, which may affect how much energy they consume. By taking into account the varied servers available in National Cloud Data Centres, Wang S. et al. [39] examined the Particle Swarm Optimization (PSO) algorithm to build a VM placement approach that reduces energy usage and enhances QoS. (NCDC).

SLA Violation based management

Big enterprise software systems, like those used by Amazon, Facebook, and others, must offer its consumers a high level of confidence regarding Quality of Service (QoS) measures including response times, high throughput, and service availability. These applications' service providers risk losing their user base and, thus, their revenue, without such guarantees. For the QoS properties, clients typically establish Service Level Agreements (SLAs) with service providers [40]. Strategies that try to reduce SLA violation may result in greater energy costs, while those that aim to increase SLA violation may result in lower energy costs. SLA violations may occur as a result of system failure, increased network load, and resource unavailability [41], [42]. SLA violations in environments that conserve energy are primarily caused by the consolidation of VMs onto a small number of servers [43]. If VM requests more resources in this case, SLA violations may result from PM's lack of resources.

A key responsibility of cloud service providers is the efficient and effective management and distribution of resources among clients. To manage the resources of the data centre in a way that is both economical and SLA-compliant, Xiong et al. [44] introduced an intelligent resource management system called SmartSLA. The SmartSLA takes into account both local and global analyses, two crucial issues. The former issue is to identify the right configuration of system resources to meet the clients' SLA. The later issue is to take decision on how resources among clients are based on current system status. This system captures relationship between system resources and performance using machine learning technique.

Future Workload based management

Static values of upper and lower thresholds on servers are inappropriate since the resource demands of applications are dynamic and unpredictable in nature.

So, there should be some methods for dynamically adjusting threshold levels in accordance with application workload patterns. Beloglazov et al. [3], [45] looked at cutting-edge methods for auto-adjustments of usage thresholds based on statistical analysis of historical data gathered throughout the lifespan of the VM. The main concept behind this suggested strategy is to change the upper utilisation threshold's value based on the variance in CPU use.

Chowdhary et al. [46] presented various new methods based on different bin packing techniques for VM placement with customised adjustments in place of the Modified Best Fit Decreasing (MBFD) algorithm for VM placement, using similar heuristics for dynamic VM consolidation. In this study, authors found a clustering technique that groups VMs according to their CPU and memory usage, as opposed to sorting them in decreasing order as in [45]. A cluster is just a list of virtual machines, and throughout the virtual machine placement process, the highest density clusters—that is, VM lists with the most virtual machines—will be given preference. In this manner, a maximum number of virtual machines (VMs) that are similar to one another in terms of CPU usage and currently assigned RAM are first placed into hosts, after which VMs from a group of second dense cluster are permitted to be installed on eligible hosts [46].

Fragmentation based management

Virtual machines need a variety of resources to run their loaded applications, including processing power, memory, bandwidth, and storage space. Only when there are adequate resources for all dimensions are VMs launched on a certain host server or PM [47]. Therefore, some PM may have unutilized resources, referred to as resource fragments that means resources are wasted. EAGLE is an energy-efficient online VM placement technique that decreases resource fragment sizes by balancing resource usage, as proposed by Li et al. in their paper [36]. EAGLE examines the posterior resource utilisation state for each potential PM during VM deployment and then chooses the best PM in accordance with the suggested model.

Exponentially weighted moving average (EWMA) was calculated by Xiao et al. [48] based on historical predicted and observed CPU load to forecast the future resource demand of VMs. In order to quantify unevenness in the multi-dimensional resource use of a server, the authors establish the concept of skewness. A periodic method that estimates the status of resource allocation based on anticipated future resource demands of VMs helps to reduce skewness or fragmentation. The skewness algorithm defines different server threshold values, including hot, cold, warm, and green computing thresholds. Various resource categories may have distinct threshold values; for instance, the CPU and memory may have hot thresholds of 90% and 80%, respectively.

Scalability based management

The data center's virtualized infrastructure is used to meet the resource demands of cloud applications. Applications are given physical resources and separate virtual machines by the cloud service provider. Poor resource usage may result from statistically dividing physical resources into virtual machines based on an application's peak demand. Under-provisioning undoubtedly leads in unhappy customers, whereas over-provisioning resources lowers profit margin [49], [50]. Without interrupting service, it is obvious that the approach is to dynamically scale resources based on workload need. Resource allocation for VMs needs to be ramped up or scaled down because the load on applications is dynamic. VM based dynamic scalability is extensively used solution to address these requirements of web applications [51]. The implementation of VM-based scaling can be accomplished in one of two ways: modifying the allocation of resources (such as processing power, memory, storage, etc.) within a VM [35]. Both of these scaling methods are known as horizontal scaling and vertical scaling, respectively.

Virtual Machines are mapped to the proper PM according to their current resource consumption until resources are available on the PM. When applications operating under a VM need more resources and none are available on the allocated PM, the VM is moved to another PM that can meet the need. Both static and dynamic VM migration are possible [52]. With a static migration, the VM is turned off and only the configuration file is transmitted from the source server to the destination server. Dynamic or live VM migration necessitates the transfer of working state and memory from the source to the destination server; as a result, it uses a lot of I/O and network traffic [53], [54], which has a big impact on the performance of applications [52], [55]. The benefit of live VM migration is that there is no interruption of the functioning of apps operating inside the VM while the transfer is taking place.

3. CONCLUSION AND FUTURE SCOPE

A technique that balances several competing performance metrics, such as energy consumption, SLA violation, and cost, is both a fascinating and difficult task, according to the literature review that was done. The work reported here is motivated to address the various research issues such as

- A plan for consolidating and placing VMs that will shorten the time needed to find the right PM or host.
- To prevent the overloading issue, a system for placing VM into PM must be established.
- To prevent performance reduction from migration, a plan must be devised to delay VM migration as long as feasible during the VM consolidation process.
- Due to the erratic nature of workload patterns in applications, a method for scaling up or scaling down of allotted resources of VM should be devised.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Niehorster, O., Krieger, A., Simon, J., & Brinkmann, A. (2011, September). Autonomic resource management with support vector machines. In Grid Computing (GRID), 2011 12th IEEE/ACM International Conference on (pp. 157-164). IEEE.

[2] Tchana, A., Tran, G. S., Broto, L., Depalma, N., & Hagimont, D. (2013). Two levels autonomic resource management in virtualized IaaS. Future Generation Computer Systems, 29(6), 1319-1332.

[3] Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. Concurrency and Computation: Practice and Experience, 24(13), 1397-1420.

[4] Tziritas, N., Xu, C. Z., Loukopoulos, T., Khan, S. U., & Yu, Z. (2013, October). Application-aware workload consolidation to minimize both energy consumption and network load in cloud environments. In 2013 42nd International Conference on Parallel Processing (pp. 449-457). IEEE.

[5] Quang-Hung, N., Nien, P. D., Nam, N. H., Tuong, N. H., & Thoai, N. (2013, March). A genetic algorithm for power-aware virtual machine allocation in private cloud. In Information and Communication Technology-EurAsia Conference (pp. 183-191). Springer, Berlin, Heidelberg.

[6] Hyser, C., McKee, B., Gardner, R., & Watson, B. J. (2007). Autonomic virtual machine placement in the data center. Hewlett Packard Laboratories, Tech. Rep. HPL-2007-189, 189.

[7] Speitkamp, B., & Bichler, M. (2010). A mathematical programming approach for server consolidation problems in virtualized data centers. IEEE Transactions on services computing, 3(4), 266-278.

[8] Chaisiri, S., Lee, B. S., & Niyato, D. (2009, December). Optimal virtual machine placement across multiple cloud providers. In Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific (pp. 103-110). IEEE.

[9] Van, H. N., Tran, F. D., & Menaud, J. M. (2009, May). Autonomic virtual resource management for service hosting platforms. In Software Engineering Challenges of Cloud Computing, 2009. CLOUD'09. ICSE Workshop on (pp. 1-8). IEEE.

[10] Quiroz, A., Kim, H., Parashar, M., Gnanasambandam, N., & Sharma, N. (2009, October). Towards autonomic workload provisioning for enterprise grids and clouds. In Grid Computing, 2009 10th IEEE/ACM International Conference on(pp. 50-57). IEEE.

[11] Yang, C. T., Cheng, H. Y., & Huang, K. L. (2011). A dynamic resource allocation model for virtual machine management on cloud. In Grid and distributed computing (pp. 581-590). Springer, Berlin, Heidelberg.

[12] Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., & Zagorodnov, D. (2009, May). The eucalyptus open-source cloud-computing system. In Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on (pp. 124-131). IEEE.

[13] Younge, A. J., Von Laszewski, G., Wang, L., Lopez-Alarcon, S., & Carithers, W. (2010, August). Efficient resource management for cloud computing environments. In Green Computing Conference, 2010 International (pp. 357-364). IEEE.

[14] Roytman, Alan, Aman Kansal, Sriram Govindan, Jie Liu, and Suman Nath. "Algorithm design for performance

aware VM consolidation." Technical report, Microsoft Research, 2013. Number MSR-TR-2013-28 (2013).

[15] Carrera, D., Steinder, M., Whalley, I., Torres, J., & Ayguadé, E. (2008, April). Utility-based placement of dynamic web applications with fairness goals. In Network Operations and Management Symposium, 2008. NOMS 2008. IEEE (pp. 9-16). IEEE.

[16] Goudarzi, H., & Pedram, M. (2012, June). Energy-efficient virtual machine replication and placement in a cloud computing system. In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on (pp. 750-757). IEEE.

[17] Shouraboura, C., & Bleher, P. (2011). Placement of applications in computing clouds using Voronoi diagrams. Journal of Internet Services and Applications, 2(3), 229-241.

[18] Karve, A., Kimbrel, T., Pacifici, G., Spreitzer, M., Steinder, M., Sviridenko, M., & Tantawi, A. (2006, May). Dynamic placement for clustered web applications. In Proceedings of the 15th international conference on World Wide Web (pp. 595-604). ACM.

[19] Jayasinghe, D., Pu, C., Eilam, T., Steinder, M., Whally, I., & Snible, E. (2011, July). Improving performance and availability of services hosted on iaas clouds with structural constraint-aware virtual machine placement. In 2011 IEEE International Conference on Services Computing (pp. 72-79). IEEE.

[20] Jammal, M., Kanso, A., & Shami, A. (2015, June). High availability-aware optimization digest for applications deployment in cloud. In ICC (pp. 6822-6828).

[21] Meng, X., Pappas, V., & Zhang, L. (2010, March). Improving the scalability of data center networks with traffic-aware virtual machine placement. In INFOCOM, 2010 Proceedings IEEE(pp. 1-9). IEEE.

[22] Biran, O., Corradi, A., Fanelli, M., Foschini, L., Nus, A., Raz, D., & Silvera, E. (2012, May). A stable network-aware vm placement for cloud systems. In Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on (pp. 498-506). IEEE.

[23] Vu, H. T., & Hwang, S. (2014). A traffic and power-aware algorithm for virtual machine placement in cloud data center. International Journal of Grid & Distributed Computing, 7(1), 350-355.

[24] Dias, D. S., & Costa, L. H. M. (2012, December). Online traffic-aware virtual machine placement in data center networks. In Global Information Infrastructure and Networking Symposium (GIIS), 2012 (pp. 1-8). IEEE.

[25] Dupont, C., Schulze, T., Giuliani, G., Somov, A., & Hermenier, F. (2012, May). An energy aware framework for virtual machine placement in cloud federated data centres. In Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy), 2012 Third International Conference on (pp. 1-10). IEEE.

[26] Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future generation computer systems, 28(5), 755-768.

[27] Kaplan, J. M., Forrest, W., & Kindler, N. (2008). Revolutionizing data center energy efficiency. Technical report, McKinsey & Company.

[28] Bilal, K. (2014). Analysis and Characterization of Cloud Based Data Center Architectures for Performance, Robustness, Energy Efficiency, and Thermal Uniformity (Doctoral dissertation, North Dakota State University).

[29] Khosravi, A., Garg, S. K., & Buyya, R. (2013, August). Energy and carbon-efficient placement of virtual machines in distributed cloud data centers. In European Conference on Parallel Processing (pp. 317-328). Springer, Berlin, Heidelberg.

[30] Kolpe, T., Zhai, A., & Sapatnekar, S. S. (2011, March). Enabling improved power management in multicore processors through clustered DVFS. In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011 (pp. 1-6). IEEE.

[31] Guérout, T., Monteil, T., Da Costa, G., Calheiros, R. N., Buyya, R., & Alexandru, M. (2013). Energy-aware simulation with DVFS. Simulation Modelling Practice and Theory, 39, 76-91.

[32] Feng, W. C., Feng, X., & Ge, R. (2008). Green supercomputing comes of age. IT professional, (1), 17-23.

[33] Wu, C. M., Chang, R. S., & Chan, H. Y. (2014). A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. Future Generation Computer Systems, 37, 141-147.

[34] Calheiros, R. N., & Buyya, R. (2014, December). Energy-efficient scheduling of urgent bag-of-tasks applications in clouds through DVFS. In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on (pp. 342-349). IEEE.

[35] Wang, W., Chen, H., & Chen, X. (2012, September). An availability-aware virtual machine placement approach for dynamic scaling of cloud applications. In Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on (pp. 509-516). IEEE.

[36] Li, X., Qian, Z., Lu, S., & Wu, J. (2013). Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center. Mathematical and Computer Modelling, 58(5-6), 1222-1235.

[37] Minas, L., & Ellison, B. (2009). Energy efficiency for information technology: How to reduce power consumption in servers and data centers. Intel Press..
[38] Mastroianni, C., Meo, M., & Papuzzo, G. (2013). Probabilistic consolidation of virtual machines in self-organizing cloud data centers. IEEE Transactions on Cloud Computing, 1(2), 215-228.
[39] Wang, S., Zhou, A., Hsu, C. H., Xiao, X., & Yang, F. (2016). Provision of data-intensive services through energy-and qos-aware virtual machine placement in national cloud data centers. IEEE Trans. Emerging Topics Comput., 4(2), 290-300.
[40] Roy, N., Dubey, A., & Gokhale, A. (2011, July). Efficient autoscaling in the cloud using predictive models for workload forecasting. In Cloud Computing (CLOUD), 2011 IEEE International Conference on (pp. 500-507). IEEE.
[41] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems, 25(6), 599-616.
[42] Buyya, R., Garg, S. K., & Calheiros, R. N. (2011, December). SLA-oriented resource provisioning for cloud computing: Challenges, architecture, and solutions. In Cloud and Service Computing (CSC), 2011 International Conference on (pp. 1-10). IEEE.
[43] Mustafa, S., Nazir, B., Hayat, A., & Madani, S. A. (2015). Resource management in cloud computing: Taxonomy, prospects, and challenges. Computers & Electrical Engineering, 47, 186-203.
[44] Xiong, P., Chi, Y., Zhu, S., Moon, H. J., Pu, C., & Hacigümüş, H. (2011, April). Intelligent management of virtualized resources for database systems in cloud environment. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on (pp. 87-98). IEEE.
[45] Beloglazov, A., Buyya, R., Lee, Y. C., & Zomaya, A. (2011). A taxonomy and survey of energy-efficient data centers and cloud computing systems. In Advances in computers (Vol. 82, pp. 47-111). Elsevier.
[46] Chowdhury, M. R., Mahmud, M. R., & Rahman, R. M. (2015). Implementation and performance analysis of various VM placement strategies in CloudSim. Journal of Cloud Computing, 4(1), 20.
[47] Stillwell, M., Vivien, F., & Casanova, H. (2012, May). Virtual machine resource allocation for service hosting on heterogeneous distributed platforms. In IPDPS 2012. IEEE, x-pays= US.
[48] Xiao, Z., Song, W., & Chen, Q. (2013). Dynamic resource allocation using virtual machines for cloud computing environment. IEEE Trans. Parallel Distrib. Syst., 24(6), 1107-1117.
[49] Chieu, T. C., & Chan, H. (2011, October). Dynamic resource allocation via distributed decisions in cloud environment. In e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on (pp. 125-130). IEEE.
[50] Tang, C., Steinder, M., Spreitzer, M., & Pacifici, G. (2007, May). A scalable application placement controller for enterprise data centers. In Proceedings of the 16th international conference on World Wide Web (pp. 331-340). ACM.
[51] Knauth, T., & Fetzer, C. (2011, July). Scaling non-elastic applications using virtual machines. In Cloud Computing (CLOUD), 2011 IEEE International Conference on (pp. 468-475). IEEE.
[52] Zhuang, Z., & Guo, C. (2013, December). Ocpa: An algorithm for fast and effective virtual machine placement and assignment in large scale cloud environments. In Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on (pp. 254-259). IEEE.
[53] Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing—The business perspective. Decision support systems, 51(1), 176-189.
[54] Liu, H., Jin, H., Liao, X., Yu, C., & Xu, C. Z. (2011). Live virtual machine migration via asynchronous replication and state synchronization. IEEE Transactions on Parallel and Distributed Systems, 22(12), 1986-1999.
[55] Shen, Z., Subbiah, S., Gu, X., & Wilkes, J. (2011, October). Cloudscale: elastic resource scaling for multi-tenant cloud systems. In Proceedings of the 2nd ACM Symposium on Cloud Computing (p. 5). ACM.