

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015,
Nagpur, INDIA

A Survey of Virtual Machine Placement Techniques in a Cloud Data Center

Zoha Usmani^a, Shailendra Singh^b, Senior Member IEEE, a*

^a*PG Scholar, Department of CEA, NITTTR, Bhopal 462002, India*

^c*Professor & Head, Department of CEA, NITTTR, Bhopal 462002, India*

Abstract

Energy consumption of massive-scale cloud data centers is increasing unacceptably. There is a need to improve the energy efficiency of such data centers using Server Consolidation which aims at minimizing the number of Active Physical Machines (APMs) in a data center. Effective VM placement and migration techniques act as a key to optimum consolidation. Many of the recently proposed techniques realize dynamic consolidation while optimizing the VM placement. This paper presents a comprehensive study of the state-of-the-art VM placement and consolidation techniques used in green cloud which focus on improving the energy efficiency. A detailed comparison is presented, revealing pitfalls and suggesting improvisation methods along this direction.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: Cloud Computing; Data Center; Green Computing; Server Consolidation; Virtual Machine Placement

1. Introduction

The highly flexible IaaS (Infrastructure as a Service) model of cloud computing has been rapidly adopted by many enterprises as an alternative to raise their Total Cost of Ownership (TCO), providing optimal utilization of resources and money. Today a lot of companies like Amazon with its EC2 (Elastic Compute Cloud) are switching towards greening their data centers, i.e. trying to use least numbers of actively running servers. This calls for

* Corresponding author. Tel.: +91-997-705-3137.

E-mail address: zohausmani05@gmail.com

appropriate solutions like Virtualization which acts as a backbone behind the cloud computing technology. It enables sharing of computer hardware by partitioning the computational resources. A small software program, i.e. the Hypervisor or Virtual Machine Monitor (VMM) manages and controls all VM related operations. Live migration of virtual machines facilitates load balancing and consolidation. In a datacenter, often many services only need a small portion of the total available resources. This can lead to a scenario in which several virtualized servers operate and consume an increased amount of space and resources than expected and that it cannot be justified by their workload. This problem is referred to as *server sprawl*.

To prevent such wastage of resources, multiple virtual machines are packed on fewest possible physical servers and rest of the extraneous servers are turned down to sleep mode (low-power state). This approach can avoid server sprawl and is termed as *Server Consolidation*. The reduction in the number of servers has a noticeable benefit on data centers by improving system availability, reducing infrastructure complexity and of course saving energy and money. VM consolidation can achieve its goal of increasing the amount of suspended servers, but in an IAAS environment, there may arise some problems which might influence the energy efficiency of cloud.

- The first problem is the **trade-off between the performance, energy consumption & the resource utilization of running VMs**. Since they need to compete for the resources provided by their corresponding PM and share them too. The resources being PM's—CPU, main memory and I/O capabilities, the server suspensions and performance degradations can increase the execution time which in turns decreases the energy savings.
- Another concern is the **overhead caused by continuous live migration of VMs during consolidation**, due to the continuously changing resource demands. Whenever a VM is migrated, its CPU state, main memory, storage & network connections are also taken care of.
- The next concern lies in **prediction of energy consumed by the server**. This is quite complicated and requires energy-performance profiling techniques, but these are unacceptable due to their respective overheads.
- In particular, cloud data center resources (e.g. CPU, memory, network bandwidth and storage) need to be allocated with equal focus on reduction of energy usage as on satisfaction of Quality of Service (QoS) requirements specified by users via Service Level Agreements (SLAs).
- Final concern is that bringing about an efficient server consolidation is quite a complex NP-Hard problem.

Keeping these challenging aspects in mind many effective approaches of server consolidation have been introduced till date. These state-of-the-art techniques address some important issues like physical resource heterogeneity or dynamic nature of virtual machines and workloads. In this paper, we discuss the VM scheduling techniques which aim at server consolidation mainly focusing on the discussion of VM placement algorithms used by these scheduling approaches. The virtual machine scheduling mechanism is used to consolidate a data center, i.e. to reduce the number of active physical machines. As discussed earlier, dynamic server consolidation can improve the energy efficiency by optimum utilization of available resources.

1.1. Server Consolidation Steps

To solve the complex problem of dynamic server consolidation and to provide decentralization, it has been divided into four main events or steps [19] discussed as follows:

- **Host Overload Detection:** **The scheduling technique must set a threshold limit in order to decide when a certain host/server is over-utilized. This limit can be termed as 'Hot Threshold' and when this limit is crossed, some of this host's VMs need to be migrated to other hosts,**²⁷
- **Host Under-load Detection:** **If a certain server is under-utilized, i.e. it has reached below the 'Cold Threshold' (a scenario just opposite to host overload) the aim of server consolidation is to identify that server and migrate all of its virtual machines to other active hosts,**¹³. Thus the under-utilized server is freed up & it can be switched to sleep/idle mode to save power.
- **VM Selection and Migration:** **Appropriate candidates (VMs) are selected either from overloaded or under-loaded host for migration.**
- **VM Placement:** **The VM(s) selected in previous step is then placed on some other physical machine according to a suitable VM to PM mapping criteria,**²⁹.

With the help of live migration of VMs, Server Consolidation aims at achieving—least possible number of Active physical machines, packing these Active PMs with VMs as tightly as possible to increase energy efficiency

and switching the non active PMs to a power saving mode.

In this paper, we are focusing only on the fourth sub-problem of server consolidation—VM Placement problem. The VM placement problem may appear to be simple at a glance, but a closer examination unmasks its complication. Here we deeply review the VM placement problem and various approaches which are introduced to realize it. The remainder of this paper is organized as follows: In Section 2, we discuss the related work which includes detailed study of methodologies used in existing VM placement algorithms. We then give an exhaustive summary of these techniques in section 3. Finally we give our concluding remarks in section 4.

2. Related Work

There has been a wide range of contribution in the “greening” of cloud data centers. An interesting survey in²⁶ provides an understanding and need of power-aware resource provisioning. Notable amount of studies have focused on such kind of data center management. Amongst which there exist a number of methodologies to address the VM placement and consolidation problem. In²² the existing methodologies for VM placement, server consolidation and load balancing were studied to uncover their anomalies and their causations, similarly²¹ discusses the open challenges in dynamic resource allocation. Our study draws upon the advances in VM placement optimization strategies for desirable consolidation. Various VM placement heuristics are compared and upgraded in³³. The work in¹⁴ broadly categorized VM placement into two methods, namely Direct placement and Migration-based placement, based on the fact that the time taken to complete a job depends upon the type of VM placement. The performance impact of VM placement and resource contention was given in¹⁵ while the researchers in²⁴, contrary to energy saving approaches like above, tried to maximize the placement ratio aiming at density placement as their main concern.

The process of consolidation introduces some trade-offs like that between delay and migration cost, Wang et al.³⁶ try to balance it by assigning a particular weight to VMs. In^{5, 11} they observed the trade-off between energy and performance. Plus in¹¹, they inferred the operating points that optimally reduce energy consumption problem as a multi-dimensional Bin Packing problem. This study proves fruitful by throwing light on the key aspects of energy-performance relationship and uncovers many research issues. Although it is hard to come up with a green computing approach which is energy-efficient, dynamic, high in performance and to strike a full balance between allocated resource with the minimum migration overhead. Most of these approaches undergo the problem of additional migration overhead which needs to be taken care of. In the rest on the paper, the VM placement algorithms are elaborated in support of the existing methodologies used in existing literature.

1.2. Virtual Machine Placement Algorithms

Virtual Machine placement is the process of selecting the most suitable Physical Machine (PM) for a given Virtual Machine (VM). So a **VM placement algorithm aims at determining the most optimal VM to PM mapping whether it is an initial VM placement or a VM migration for placement re-optimization.** The placement technique in VM consolidation can have one of the two major goals—one is power saving and other is delivering QoS. The type of VM placement approach varies from a cloud service provider to another. There is a clear conflict between these two goals. It is noteworthy here that we are dealing only with the power-based approaches which use Dynamic VM placement algorithms (the static ones being obsolete).

2.1.1 Classification of VM Placement Algorithms

Depending on the *goal* of placement, a VM placement algorithm can be broadly categorized into two types:

- 1) **Power-based approach:** Aims to obtain a VM-PM mapping which results into a system that is energy-efficient with utmost resource utilization, [1, 13].
- 2) **QoS-based approach:** Aims to obtain a VM-PM mapping to ensure maximal fulfillment of quality of service requirements, [7, 27, 28].

Depending on the type of *principal approach* used to attain a desirable VM-PM mapping, VM Placement techniques are mainly classified as under:

1) Constraint Programming: It is a kind of logic programming, as a contrast to mathematical approaches, to solve complex combinatorial problem of optimal VM placement. It uses a set of constraints which can easily be extended further to involve more aspects.

Zhang *et al.*³⁰ proposed a virtual cloud resource allocation model based on constraint programming (VCRA-CP), capable of meeting goals of Quality of Service requirements and reducing the cost of resource usage. The authors took into account the performance fulfillment goals of applications and workload types.

Dupont *et al.*²⁹ introduced a flexible and extensible framework which is based on the VM Repacking Scheduling Problem (VRSP) for energy-aware resource allocation in data centers considering SLA constraints to perform VM placement. This approach allows the user to automatically derive the SLA constraints and lowers energy usage.

J. Dong *et al.*⁵ applied a few constraints such as network link capacity and Physical Machine (PM) size on scheduling of Virtual Machines (VM) via a two-staged VM scheduling algorithm. Firstly, to place VMs, they joined the Best Fit heuristic of Bin Packing with min-cut hierarchical clustering. This leads to minimum number of active PMs and further avoids network congestion by MLU (Maximum Link Utilization) optimization, which is achieved by modeling network traffic as a QAP (Quadratic Assignment Problem). Secondly, the allocated VMs are re-optimized.

2) Bin Packing: The classical problem of Bin packing consists of a series of items having sizes specified in the interval (0, 1] which need to be packed into least possible number of bins with capacity one. To model this problem as a resource allocation algorithm, we consider each item as a Virtual Machine (VM) to be tightly packed in minimum number of bins, each considered as a Physical Machine (PM). The bin packing problem is NP hard. The quality of a polynomial time approximation algorithm, A is measured by its approximation ratio, R (A) to optimal algorithm, OPT:

$$R(A) = \lim_{n \rightarrow \infty} \sup_{OPT(L)=n} \frac{A(L)}{OPT(L)}$$

where, A (L) is the number of bins used under the algorithm A, OPT (L) is the number of bins used under the optimal algorithm OPT and L is the list of input sequence. In this section we throw some light on the existing VM scheduling techniques which aim at improving server consolidation using Bin Packing approach.

W. Song *et al.*¹ formulated a dynamic resource allocation algorithm based on Bin packing which optimizes the number of actively running servers. They designed a slight variation of the Relaxed Online Bin Packing algorithm⁶ and named it as VISBP (Variable Item Size Bin Packing). They implemented it using extensive trace-driven simulation and also compare it with three well known server consolidation algorithms: the Black Box & Gray box algorithm¹², the Vector Dot algorithm¹⁶ and the Offline-Bin Packing algorithm⁷. The core of VISBP is its ability to handle the change in size of an item (VM) at runtime. This “change” operation supports an on demand, dynamic resource allocation. VISBP excels in load balancing and hot-spot detection but it violates service level agreements to an extent and there is need to improve the VM to PM ratio.

Y. Zhang *et al.*⁴, addressed the problem that the existing Bin packing heuristics, whether single dimensional or multi-dimensional, do not dig much into the resource requirement heterogeneity of VMs. So they proposed heterogeneity aware algorithms like DRR-FFD (Dominant-Residual Resource aware FFD) and its variations. With a little struggle in clustering VMs based on their dominant resources, the proposed algorithms achieve a performance similar to multi-dimensional algorithms.

C. Lin *et al.*¹⁸, proposed two unique algorithms for energy-effective virtual machine provisioning and consolidation, first is Dynamic Round-Robin (DRR) and second is Hybrid which is a fusion of Dynamic Round-Robin and First-Fit. They also shared some implementation issues like ‘retirement threshold’ of a physical machine. The proposed algorithms reduce the average power-consumption.

3) Stochastic Integer Programming: In contrast to logical approach, this is a mathematical optimization technique in which the future demands are uncertain. They make use of estimation models using probability distributions of the concerned data. Here, the future demand of a VM or an application is unknown and therefore, some VM placement techniques use this approach to predict the suitable VM-PM mapping.

N. Bobroff *et al.*⁷, devised a dynamic server consolidation and migration algorithm which decrements the SLA violation rate and reduces the capacity demands of servers, thereby curtailing data center running costs. The algorithm operates in three major steps- (1) Measuring historical data (2) Forecasting the future demand (3)

Remapping VM to PM, and therefore it is referred to as Measure-Forecast-Remap (MFR) algorithm.

Speitkamp *et al.*³¹ provided a mathematical formulation of the NP-Hard Optimization problem for server consolidation making use of an LP-relaxation-based heuristic and historical workload analysis. They extended the SSAPv and DSAP decision models explained in³², applying a number of constraints to minimize the servers' operational and other costs. This capacity planning approach uses an optimization model along with a data preprocessing approach to achieve optimal placement.

M. Chen *et al.*⁹ assigned virtual machines to physical machines with probability of server load exceeding its capacity being ' p '. This 'Effective VM Sizing' estimates the VM resource demand as an aggregation of intrinsic demand and correlation-aware demand. This is a consequence of statistical multiplexing in data centers where the same server is packed with multiple VMs.

4) Genetic Algorithm: Being a part of evolutionary computation, it performs natural selection of suitable solution from all possible solutions. This heuristic can be called as **bin packing extended with additional constraints**.

Mi *et al.* [34] propose a **Genetic Algorithm Based Approach (GABA) which follows an adaptive self-reconfiguration of VM reallocation on heterogeneous PMs**. It can search for optimal solutions online. To catch up with the changing workloads, request forecasting module is used. GABA results in conservation of power and deals with multi-objective optimization.

Ferdaus *et al.*²⁰ **model the problem of VM placement** as an NP-Hard Multi-Dimensional Vector Packing Problem (mDVPP) focusing on **balancing the cloud resource utilization, making use of the ACO (Ant Colony Optimization) metaheuristic**. This is an effective approach where computation time is also remarkably lesser.

Gao *et al.*²³ **minimized the power consumption and resource wastage of VM placement problem** using a modification of **Ant Colony System (ACS) algorithm**. The residual resources were effectively balanced along different dimensions on the servers. This combinatorial problem is modeled as a multi-objective algorithm named VMPACS.

3. A Summary of VM Placement Techniques

Table 1 summarizes the already discussed algorithms and techniques in the previous section by comparing their basic attributes like—the standard VM placement technique on which they are based (out of the four *principal approaches* mentioned above), the number of resources used (like CPU, memory and bandwidth), their achieved goals, future enhancements and lastly, those well known algorithms which they outperform. The progress of VM placement optimization techniques can be clearly observed in the table.

The state-of-the-art VM placement algorithms range from static to dynamic (although we have considered only the dynamic ones). Further they are extended to adaptive algorithms [1, 3, 5, 34, etc] which have the ability to mold themselves according to the change in workload and demand. One can easily observe that to achieve an optimal server consolidation, different approaches focus on different aspects of optimization scenario. The *constraint-based* approach can reduce the number of APMs but the prolonged search time is a major drawback. The *bin packing-based* algorithms can substantially decrease the required amount of active servers (APMs) and so reduce the required energy usage but they sometimes result in extremely tight packing of VMs on PMs. *Stochastic* techniques should be an option to use when there is more uncertainty in parameters which can affect the costs. Otherwise the forecasting techniques require seeking of large search spaces. Lastly, *Genetic algorithms* suffer from the drawback of much higher computation time. Collectively looking onto all these techniques, we observe that the VM resizing techniques are highly efficient in terms of cost reduction; however the resizing of VMs imposes extra overhead depending upon the 'resize criterion' as in Sandpiper¹². But the strict server consolidation techniques impose an extra overhead of migration. A power based consolidation performs the maximum possible use of available resources thereby compromising on QoS and violates SLA constraints. Some algorithms only consider CPU as their primary resource and thus they should be extended to consider other important resources that cannot be sidelined.

Table 1. Comparison of State-of-the-art VM Placement Algorithms.

Algorithm	Based on	Resources considered	New Aspect	Strength	Weakness	Performance Better Than
Exact Allocation and Migration [2]	Constraint Programming	CPU	Objective functions for optimality	Reduced the number of APMs, Reduced migration cost	Slightly slow execution	Best-fit heuristic
Min-Cut hierarchical clustering [5]	Constraint Programming	CPU & network bandwidth	MLU optimization & VM reuse	Reduced energy consumption and network traffic	More VM-migration cost	BFD & Random algorithms
MFR (Measure-Forecast-Remap) [7]	Stochastic integer programming	CPU	Time interval of length ' τ '	Meets SLA targets, Reduced the number of APMs	Need for extension to multiple resources	Static algorithm
Group Packing Algorithms [8]	Stochastic bin packing	CPU & Bandwidth	Random variables to predict future Bandwidth	Reduced the number of APMs	Need for extension to multiple resources	First-fit, FFD & Harmonic algorithm
Effective VM Sizing Algorithms [9]	Stochastic integer programming	CPU, memory as a constraint	Server overflow probability ' p '	Reduced the number of APMs & O(1)-approximation	Need for extension to multiple resources	FFD algorithm
VectorDot [16]	Bin Packing	CPU, memory, network, i/o bandwidth	EVP metric, Integrated server and storage migration	Dynamic load balancing, Managing overloaded nodes	Need to follow predictive and statistical models	Best-fit, First-fit, Worst-fit & Relaxed-Best-fit heuristics
Sandpiper: Black-box & [12]	Bin Packing	CPU, memory & network	VSR (Volume to Size ratio)	Hot-spot detection & mitigation, Load balancing	VM resizing & Migration overhead	-
VISBP [1]	Bin Packing	CPU, memory & network	'change' operation (change in size of VM)	Hot-spot mitigation, Load balancing	SLA violation	Sandpiper, VectorDot, Offline BP algorithms
Enhanced FFD [3]	Bin Packing	CPU	VM reuse strategy	More energy-efficient, High system throughput	SLA violation	Greedy, Round-Robin & FFD algorithms
Heaviest First [17]	Bin Packing	CPU	Upper bound on cost of VM relocation	Reduced cost of relocation	Uses slightly more number of bins	Best-fit, First-fit
GABA [34]	Genetic Algorithm	CPU	Request forecasting and Reconfiguration searching module	Reduced the number of APMs, Improved CPU utilization	Overhead of large searching spaces	TSSP07 Approach in [35]

VMPACS [23]	Genetic Algorithm	CPU, network and storage	ACS meta- heuristic applied to multi-objective VM placement Problem	Near-optimal solution, Energy- efficient, Minimum resource wastage	High computation time	MGGA (Multi- Objective Grouping Genetic Algorithm)
AVVMC	Genetic Algorithm	CPU, memory, network i/o	Vector-algebra based capturing of resource utilization	Improved overall resource utilization	Migration and re- configuration overheads	MMVMC (Max-Min VM Consolidation)
BGM-BLA [25]	Genetic Algorithm	CPU, memory & Storage	Bucket coding and learning	Near-optimal solution, Reduced energy consumption	Bucket code is asymmetric for small problems	NSGA-II, BGM-CA (common Coding algorithm)

4. Conclusion

Server Consolidation in data centers has been an active area of research in the past few years. This survey throws light on such consolidation mainly detailing the VM placement algorithms and methods used to reach an optimal solution for this placement problem. The objective of these techniques can either be minimization of power consumption or providing QoS, both being in conflict. Ranking these algorithms or stating the best one out of the lot is not a proper suggestion because every other placement technique has some specific target, migration technique, prominent resources and influential parameters. Although these techniques may seem fine from outside, there exist some or the other kind of trade-offs when deeply surveyed. Owing to the workload-variability and continuously changing demands of applications, there is a need to constantly optimize these VM placement algorithms.

As a future enhancement, we can suggest that there can be an approach which minimizes the trade-off between energy consumption and good performance (QoS). This can be done by using a hybrid technique which uses a combination of server consolidation for energy-efficiency as well as it performs load balancing to deliver better quality of service. This is a two-staged process comprising—firstly, ‘*Green computing*’ for maximum resource utilization, this can be done using any of the above mentioned VM placement method (example: Bin Packing). Secondly we can perform ‘*Overload-Avoidance*’ for load balancing. The overloading can be avoided by continuously keeping a check on the ‘Hot Threshold’. As a future enhancement, one can contribute towards greening of data centers by implementing the above suggested hybrid technique.

References

1. Song, W., Xiao, Z., Chen, Q. and Luo, H., 2014. Adaptive resource provisioning for the cloud using online bin packing. *Computers, IEEE Transactions on*, 63(11), pp.2647-2660.
2. Ghribi, C., Hadji, M. and Zeghlache, D., 2013, May. Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms. In *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on* (pp. 671-678). IEEE.
3. Alahmadi, A., Alnowiser, A., Zhu, M.M., Che, D. and Ghodous, P., 2014, March. Enhanced First-Fit Decreasing Algorithm for Energy-Aware Job Scheduling in Cloud. In *Computational Science and Computational Intelligence (CSCI), 2014 International Conference on* (Vol. 2, pp. 69-74). IEEE.
4. Zhang, Y. and Ansari, N., 2013, December. Heterogeneity aware dominant resource assistant heuristics for virtual machine consolidation. In *Global Communications Conference (GLOBECOM), 2013 IEEE* (pp. 1297-1302). IEEE.
5. Dong, J., Wang, H. and Cheng, S., 2015. Energy-performance tradeoffs in IaaS cloud with virtual machine scheduling. *Communications, China*, 12(2), pp.155-166.
6. Gambosi, G., Postiglione, A. and Talamo, M., 2000. Algorithms for the relaxed online bin-packing model. *SIAM journal on computing*, 30(5), pp.1532-1551.
7. Bobroff, N., Kochut, A. and Beaty, K., 2007, May. Dynamic placement of virtual machines for managing sla violations. In *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on* (pp. 119-128). IEEE.

8. Wang, M., Meng, X. and Zhang, L., 2011, April. Consolidating virtual machines with dynamic bandwidth demand in data centers. In *INFOCOM, 2011 Proceedings IEEE* (pp. 71-75). IEEE.
9. Chen, M., Zhang, H., Su, Y.Y., Wang, X., Jiang, G. and Yoshihira, K., 2011, May. Effective VM sizing in virtualized data centers. In *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on* (pp. 594-601). IEEE.
10. Rieck, B., 2010. Basic Analysis of Bin-Packing Heuristics. *Publicado por Interdisciplinary Center for Scientific Computing. Heildelberg University*.
11. Srikantaiah, S., Kansal, A. and Zhao, F., 2008, December. Energy aware consolidation for cloud computing. In *Proceedings of the 2008 conference on Power aware computing and systems* (Vol. 10).
12. Wood, T., Shenoy, P., Venkataramani, A. and Yousif, M., 2009. Sandpiper: Black-box and gray-box resource management for virtual machines. *Computer Networks*, 53(17), pp.2923-2938.
13. Beloglazov, A. and Buyya, R., 2010, November. Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers. In *Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science* (Vol. 4). ACM.
14. Li, K., Zheng, H. and Wu, J., 2013, November. Migration-based virtual machine placement in cloud systems. In *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on* (pp. 83-90). IEEE.
15. Paul, I., Yalamanchili, S. and John, L.K., 2012, December. Performance impact of virtual machine placement in a datacenter. In *Performance Computing and Communications Conference (IPCCC), 2012 IEEE 31st International* (pp. 424-431). IEEE.
16. Singh, A., Korupolu, M. and Mohapatra, D., 2008, November. Server-storage virtualization: integration and load balancing in data centers. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing* (p. 53). IEEE Press.
17. Singh, A., Korupolu, M. and Mohapatra, D., 2008, November. Server-storage virtualization: integration and load balancing in data centers. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing* (p. 53). IEEE Press.
18. Lin, C.C., Liu, P. and Wu, J.J., 2011, December. Energy-efficient virtual machine provision algorithms for cloud systems. In *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on* (pp. 81-88). IEEE.
19. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers
20. Ferdous, M.H., Murshed, M., Calheiros, R.N. and Buyya, R., 2014. Virtual machine consolidation in cloud data centers using ACO metaheuristic. In *Euro-Par 2014 Parallel Processing* (pp. 306-317). Springer International Publishing.
21. Buyya, R., Beloglazov, A. and Abawajy, J., 2010. Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges. *arXiv preprint arXiv:1006.0308*.
22. Mishra, M. and Sahoo, A., 2011, July. On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach. In *Cloud Computing (CLOUD), 2011 IEEE International Conference on* (pp. 275-282). IEEE.
23. Gao, Y., Guan, H., Qi, Z., Hou, Y. and Liu, L., 2013. A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. *Journal of Computer and System Sciences*, 79(8), pp.1230-1242.
24. Camati, R.S., Calsavara, A. and Lima Jr, L., 2014. Solving the virtual machine placement problem as a multiple multidimensional knapsack problem. *ICN 2014*, p.264.
25. Tao, F., Li, C., Liao, T. and Laili, Y., 2015. BGM-BLA: a new algorithm for dynamic migration of virtual machines in cloud computing.
26. Beloglazov, A., Buyya, R., Lee, Y.C. and Zomaya, A., 2011. A taxonomy and survey of energy-efficient data centers and cloud computing systems. *Advances in computers*, 82(2), pp.47-111.
27. Beloglazov, A. and Buyya, R., 2013. Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints. *Parallel and Distributed Systems, IEEE Transactions on*, 24(7), pp.1366-1379.
28. Beloglazov, A. and Buyya, R., 2010, May. 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* (pp. 826-831). IEEE Computer Society.
29. Dupont, C., Giuliani, G., Hermenier, F., Schulze, T. and Somov, A., 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.
30. Zhang, L., Zhuang, Y. and Zhu, W., 2013. Constraint Programming based Virtual Cloud Resources Allocation Model. *International Journal of Hybrid Information Technology*, 6(6), pp.333-344.
31. Speitkamp, B. and Bichler, M., 2010. A mathematical programming approach for server consolidation problems in virtualized data centers. *Services Computing, IEEE Transactions on*, 3(4), pp.266-278.
32. Bichler, M., Setzer, T. and Speitkamp, B., 2006. Capacity planning for virtualized servers. In *Workshop on Information Technologies and Systems (WITS), Milwaukee, Wisconsin, USA*.
33. Chowdhury, M.R., Mahmud, M.R. and Rahman, R.M., 2015, June. Study and performance analysis of various VM placement strategies. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACI International Conference on* (pp. 1-6). IEEE.
34. Mi, H., Wang, H., Yin, G., Zhou, Y., Shi, D. and Yuan, L., 2010, July. Online self-reconfiguration with performance guarantee for energy efficient large-scale cloud computing data centers. In *Services Computing (SCC), 2010 IEEE International Conference on* (pp. 514-521). IEEE.
35. Tang, C., Steinder, M., Spreitzer, M. and 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.
36. Wang, X., Chen, X., Yuen, C., Wu, W. and Wang, W., 2014, December. To migrate or to wait: Delay-cost tradeoff for cloud data centers. In *Global Communications Conference (GLOBECOM), 2014 IEEE* (pp. 2314-2319). IEEE.