*Asia-Pacific J.Mgmt. Tech. Volume 3(1) 01- 12*
https://doi.org/10.46977/apjmt. 2022v03i01.001

AJMT

**Asia-Pacific Journal of Management and Technology**

Online ISSN: 2652-6840

journal homepage www.ajmt.org

Review Article

# An Overview of the State-of-the-art Virtual Machine Placement Algorithms for Green Cloud Data Centres

Anindya Bose*, Sanjay Nag

*Department of Computational Science, Brainware University, West Bengal 700125, Kolkata, India*

*Correspondence E-mail:* anindyabose04@gmail.com*, dsn.cs@brainwareuniversity.ac.in

**Abstract**

Increased energy consumption in Cloud Data Centres (CDCs) increases the carbon footprint. Efficiency of the data centres thus needs to be improved through server consolidation using effective virtual machine (VM) placement and migration techniques and minimizing the number of active physical machines (PMs). One of the problems is how to operationally allocate the VMs to PMs. These allocations have both operational costs and energy consumption issues. To achieve the aim of 'Green Computing' a number of state-of-the-art machine learning algorithms have been proposed for the VM placement. The authors of this paper have provided a detailed discussion and comparison of some of the current research works on energy efficiency. and cons of each of these techniques have been discussed. Some future research prospects in this field have also been mentioned at the end.

***Keywords:*** Cloud Data Centres; Virtual machine; Physical machine; Green computing; Server consolidation.

## Introduction

Since the last decade, rapid advancements in modern computational power-driven applications, as well as the shift to cloud computing, have resulted in large-scale virtualised data centres. Cloud Data Centres (CDCs) usually consist of thousands of servers that cater to numerous web-based applications. Resources are thus needed to be virtualised, which is a technique to abstract low-level server hardware resources. Whereas, upper-level services that are provided on demand are encapsulated in virtual machines (VM). The enormous power consumption of these data centres leads to huge carbon-dioxide emissions and high operational costs. One of the ways to reduce power consumption is to use dynamic consolidation of VMs using live migration (Jing, Ali & She, 2103). Aggressive consolidation, however, may lead to performance degradation, which may in turn lead to a fall in quality of service (Beloglazov & Buyya, 2012). Optimisation of VM is necessary due to the variation in workload imposed by the applications. Hence, dynamic consolidation of VM in green clouds is necessary with a focus on VM placement and consolidation techniques. This should be implemented while ensuring adherence to the Service Level Agreement (SLA).

Big companies like Amazon are moving towards green computing by using an optimum number of actively running servers. This calls for virtualization, which facilitates the sharing of hardware by segregating the computational resources. Many applications need only a small amount of available resources, resulting in wastage of both space and resources. To overcome this problem, multiple VMs are packed on minimal physical servers dynamically. The rest of the external servers are kept in sleep mode. This is known as server consolidation, which not only saves energy but also improves system availability and reduces infrastructure complexity. The steps involved in server consolidation are:

1. Detection of host-overload.
2. Detection of host-underload.

*Received 11 May 2022; Received in revised form 23 May 2022; Accepted 1 June 2022*

3. VM selection and migration.
4. VM placement.

There is a utilization threshold for a server, which if crossed, means the VM is migrated to other hosts. On the other hand, if there is an underutilization of a server, its VMs are migrated to active hosts.

When implementing VM placement algorithms, the PMs are usually divided into two groups – the PMs that satisfy some predefined criteria and those that do not. The PMs that meet the criteria are ordered and the VMs are placed in them till all the VMs are exhausted or the number of legitimate PMs is all used up. Many studies focus on ordering heuristics related to online bin-packing, while others consider specific attributes such as CPU usage, memory usage, and so on (Mills, Filliben, & Dabrowski, 2011).

**Problems with Virtual Machine Placement**

VM consolidation and placement may give rise to certain problems that affect the energy-efficiency of Cloud Data centres (CDC) (Usmani & Singh, 2016).

- There is competition for resources such as the CPU, memory, and I/O capabilities of the PMs. Suspension of the servers and performance degradation lead to a reduction in energy efficiency.
- Continuous change in resource requirements leads to continuous live migration, which causes several overheads.
- The energy-consumption profile of the servers is needed to perform VM placement. Prediction of energy consumption is a complex technique that causes further overhead.
- The resources of the CDC need to be allocated such that a reduction in energy consumption does not affect the QoS.
- Server consolidation is a NP-hard problem.

The prime focus of this paper is to provide an overview of the VM placement strategies that have been adopted over the years. Despite the fact that several research works provide numerous in-depth solutions, a closer examination of these techniques reveals several flaws. rest of the paper is arranged as follows: In Section 2, the authors discussed several approaches to VM placement and server consolidation; in Section 3, the pros and cons of the different techniques along with their proficiency were discussed; in Section 4, the conclusion of the overview and outline of some future research prospects.

**Review of Literature:**

There are several researches works that propose the implementation of Green CDC. A number of techniques exist that address the VM placement and server consolidation issues. Mishra and Sahoo (2011) studied the existing methods to discover the anomalies and the associated cause, while Buyya, Beloglazov and Abawajy, (2010) discovered the challenges related to dynamic resource allocation. Chowdhury, Mahmud and Rahman (2015) compared the different VM placement heuristics. Li, Zheng and Wu (2013) classified the VM placement techniques into 'direct placement' and 'migration-based placement'. Consolidation process often involves trade-offs between the cost of migration and delay - Wang *et al*. (2014) addressed this issue by assigning weights to the VMs. One work addressed the energy performance and concluded the operating points that reduced the energy consumption as a multi-dimensional bin-packing problem.

It is difficult to find a robust Green-solution that is simultaneously energy-efficient, dynamic, does not violate the SLA, and does not affect the QoS. Most of the VM placement algorithms suffer from one type of overhead or the other. Rest of this section discusses the applications of some of the common algorithms of VM placement that had been applied by several researchers.

**Virtual Machine Placement Algorithms**

In the cloud management framework, one of the most important activities is the placing of VM over physical machines (PM). Maximum resource utilization is possible if VMs can be placed over an optimal set of PMs of fixed capacity. Users, on the other hand, do not want any degradation in performance.

Virtualization overhead, which depends on the type of application, might be significant, leading to performance compromise. Thus, to ensure that the SLA is maintained, the overhead needs to be defined apriori when considering the VM placement, which is dynamic in nature, i.e., new VMs of variable requirements arrive while existing ones leave. Hence, after the initial placement, the scheduling decision needs to be updated periodically. VM placement algorithms can be classified into two major groups:

- The power-based approach has the objective of energy-efficient VM-PM mapping with optimal resource utilization.
- The QoS-based approach has the objective of VM-PM placement that assures maximum quality of service requirements.

**Application of Commonly used Algorithms for VM Placement**

VM placement algorithms find the optimal solution in an iterative manner. Some of the commonly used ones are:

*A. Best-Fit Decreasing (BFD)*

This algorithm is used for solving bin packing problems. It works on the principal of packing heterogeneous VMs onto a limited number of PMs. VMs are sorted in descending size order and inserted on the first PM of the appropriate size.

Energy and Carbon-Efficient (ECE) architecture proposed by Khosravi, Garg and Buyya (2013) not only minimized carbon footprints but also ensured QoS. They had taken into account distributed data centres with different carbon footprints and power usage effectiveness (PUE). ECE places the request for VM placement at the most suitable data centre site and physical server. ECE not only keeps information about each data centre site but also keeps information related to the parameters such as PUE, carbon footprint rate, utilization, energy consumption, etc. of the physical servers. This information is used to perform VM placement in a cloud environment. Carbon footprint was reduced by 10% and 45% compared to other carbon-efficient and non-carbon efficient heuristics, respectively. However, SLA violations for low system load were much higher with ECE.

Energy and performance efficient resource management can be applied for dynamic VM consolidation. Ratio of the cost of an offline algorithm to an online deterministic algorithm for single VM consolidation was computed. An optimal online deterministic algorithm based on local regression combined with an MMT VM selection policy was proposed for the dynamic migration of multiple VMs to multiple hosts (Beloglazov & Buyya, 2012). The proposed algorithm was evaluated using simulation on a large-scale setup.

OpenStack is the most widely used cloud management tool. It has a module to determine when to migrate a VM and for selecting a suitable host and for VM placement using the Modified Best-Fit Decreasing (MBFD) algorithm. MBFD not only increases energy consumption, it also violates SLA. Thus, though it was originally designed to minimize VM migration, MBFD actually increases it. Moges and Abebe (2019) designed a VM placement algorithm that is based on the bin-packing heuristic as well as increasing the power efficiency of the server. They also developed Medium Fit (MF), which is a new bin-packing heuristic to reduce SLA violation. Performance of the algorithms was evaluated in homogeneous, heterogeneous, and default scenarios. Energy consumption improved by 67%; VM migration and SLA violation improved by 46% and 78% respectively.

*B. Ant colony Optimization (ACO)*

It is the most popular optimization technique to find the optimal path through the graph. Ant agents locally update the pheromone values, and the most optimal path has the highest pheromone value.

Consolidation of VMs on physical servers is a multi-dimensional bin-packing (MDBP) problem that is NP-hard in nature. Selection of appropriate VMs for migration due to overload or underload is best performed using heuristic search techniques. Most of the current algorithms are greedy in nature, which gives rise to wastage of resources. Feller, Rilling, and Morin (2011) showed that the swarm intelligence approach provided better gains than the traditional approach. They modelled the VM placement as an instant of MDBP using ACO. Their experiment was conducted in a homogeneous environment when all physical machines have the same capacity. VMs were initially scheduled under the consideration that their resource requirements were static. As time passed, the history of resource utilization became available, and this data was used to compute the resource demand. The resource requirement for a period of one week was used as input.

Problem is thus, to map the VMs to the physical machines (PMs) which are bins where VMs can be placed. Predefined static items in each bin are CPU cycles, RAM, network bandwidth, disk capacity, etc. The resource capacities of each homogeneous host are 10000 MIPS, 24 cores, 50 GB RAM, 4 TB storage, and a 10 GBit/sec network connection. A total of 600 VMs were simulated with each needing 1000, 2000, 3000, or 5000 MIPS, 2 cores, 4 GB of RAM, and 1 GBit/sec network bandwidth. The problem was represented using the binary-integer programming (BIP) model. VMs are assigned to bins or PMs using a probabilistic decision rule. Since the objective is to maximize resource utilization, heuristic function is defined to favour items that utilize the bins better.

They compared the outcome of the ACO-based approach with that of the IBM ILOG CPLEX v12.2. The percentage improvement in energy gain by ACO over CPLEX is given in Table 1.

**Table 1:** Percentage improvement ACO-based approach with CPLEX approach – the result obtained with simulation

| Number of VMs | Energy Savings with ACO | Energy Gain % Over CPLEX |
|---|---|---|
| 100 | 5.88% | 0 |
| 200 | 4.47% | 25% |
| 300 | 3.98% | 20% |
| 400 | 3.73% | 28.5% |
| 500 | 4.18% | 13.1% |
| 600 | 3.96% | 21.1% |

The time required for placement with ACO, however, was much higher. Also, the ACO-based method is less optimised than CPLEX. The result was also obtained with the first-fit decreasing (FFD) heuristic algorithm, which is the most commonly used algorithm for bin packing, and it was found to take a larger number of hosts and hence a larger number of VMs. As FFD is a greedy algorithm, it failed to provide scalability and fault-tolerance.

The ACO approach taken by Liu *et al*. (2014) reduces the number of running physical servers one by one. VM placements with various demands for resources were optimally placed on the physical servers. The performance evaluation was done with up to 600 VMs and when compared with BFD, which is a first-fit decreasing (FFD) algorithm, the performance of the ACO-based approach was found to be better. ACO, however, suffers from slow convergence speed and easy stagnation. This problem was addressed in another work by proposing a pheromone diffusion model, which was found to be more efficient when tested with both synthetic and realistic data (Liu *et al*., 2017).

### *C. Constraint Programming*

It is a form of logic programming that uses a set of constraints that are satisfied at the same time. More aspects can be included if constraints are extended further.

Though the data centres are adopting power conservation approaches, there are some typical constraints that need to be taken into consideration. Data centres usually cater to a number of users with variable operational and performance requirements. It is thus necessary to separate the resource management activities from the constraints. Dupont *et al*. (2012) proposed a flexible method for the consolidation of VMs in Cloud Data Centres. They decoupled the constraints and the VM placement algorithms to address the problems related to extensibility and flexibility. The optimizer in this proposed system was able to deal with Service Level Agreement (SLA), requirements of different heterogeneous data centres, and minimization of energy consumption.

A component called the Power Calculator was used in this model. When the physical and dynamic elements of a data centre are given as input, it simulates the power consumption by every element of the data centre on a real-time basis. Since this is an NP-hard problem, they bypassed calling the Power Calculator each time VM placement was made. They used a static variation of the Power Calculator instead. The branching heuristic sorted the VMs in increasing order of their energy efficiency. Each VM was placed on a server that provided the maximum energy gain. Several constraints, such as hard disk, RAM, CPU cores, GPU cores, RAID level, bandwidth, max CPU load, planned outage, etc., were taken into account. The performance of the system was tested at two separate data centres. The reduction in power consumption and carbon dioxide emissions was approximately 18%.

In another work, they proposed an extensible architecture called Plug4Green. It calculates the placement of VMs and the state of the physical servers depending on 23 SLA constraints and 2 constraints to reduce greenhouse gas emissions and power consumption. This system allows the constraints to be formulated independently. Power consumption and greenhouse emissions were reduced by 33% and 34%, respectively. Plug4Green was simulated to show that 7,500 VMs running on 1,500 servers could be placed within one minute (Dupont *et al.*, 2017).

Dong, Wang and Cheng (2015) applied constraints such as network link capacity and PM resources like CPU, memory, etc. to the scheduling of VMs in two stages. In the first stage, VM placement scheme is implemented based on the constraints, and in the second stage, VMs are migrated dynamically to optimize the migration cost, power consumption, and network performance. A Best-Fit algorithm with minimum-cut hierarchical clustering was used for static VM placement in order to optimize the number of PMs. Network congestion was avoided by implementing a Maximum Link Utilization (MLU) optimization achieved through a Quadratic Assignment Problem (QAP).

To meet the goals of QoS and to reduce the cost of resource usage. Zhang, Zhuang and Zhu (2013) proposed a constraint programming-based model for cloud resource allocation. It takes into account heterogeneous workload, and when simulated, the model showed a reduction in QoS violations as well as a reduction in the cost of resource utilization.

*Particle Swarm Optimization (PSO)*

PSO is a bio-inspired algorithm that tries to find the optimal solution. It makes no assumption about the problem to be optimized and is capable of searching large solution space.

Particle Swarm Optimization (PSO) was employed by Wang *et al*. (2013) to improve the VM placement problem of dissimilar virtual data centre. To improve the VM placement, PSO was improved by redefining parameters and operators of PSO. Position of the particles were updated using fitness first approach. The improved PSO was able to find the optimal VM placement with minimal energy consumption. Proposed model was evaluated on a simulated data centre with 1000 clusters where each cluster had 350 diverse services. The performance of modified PSO was 13% - 23% better than other placement algorithms that used greedy-based approach.

Zhao *et al*. (2014) presented a heuristic approach that combines PSO with simulated annealing (SA) to implement energy saving VM placement selection. They compared the performance of the model with random-migration approaches and found that the proposed approach had lowest incremental energy consumption.

Another algorithm that works on the same principal as PSO is Firefly algorithm. It is a meta-heuristic algorithm motivated by the flashing behaviour phenomenon and of bioluminescent communication of the fireflies. In one of the works Firefly algorithms, which is an efficient clustering algorithm, was used to improve the energy consumption by 44.39% by reducing approximately 72.34% of migration and saving 34.36% hosts (Kansal & Chana, 2016).

*D. Genetic Algorithm (GA)*

It is a search heuristic that is based on natural evaluation. When applied to a searching problem GA picks the best solution from a set of solutions. The decision-making paradigms associated with the adaptive allocation of the VMs to PMs are:

- When to allocate.
- Which VM to relocate.
- Which PM to shut off.
- Where to place the VM that is to be relocated.

Mosa and Paton (2016) proposed a genetic algorithm approach to compute the CPU utilization of PM, based on the utilization of VM. Energy consumption and SLA violation were reduced by 5% and 36% respectively. Average number of PM shut off and VM migration were 5% and 16% respectively compared to the heuristic based algorithms.

Mosa and Paton (2016) proposed a GA based approach to adaptively self-configure the VMs in data centres with heterogeneous nodes. Minimal number of VMs required in each application and their physical locations are established dynamically. This allows the data centres to self-organize without categorical specification. The model was tested in a simulated environment with 300 PMs. This approach successfully switched off redundant PMs and increased average resource consumption. Energy consumption was thus reduced.

*E. Stochastic Integer Programming*

It is a mathematical optimization method in which the resource demands of the future are not known with certainty. Probability distribution is used to build the prediction model. This method had been used in scenarios where future demand for VM is uncertain. Thus, this method is used for predicting the suitable VM-PM mapping.

Measure Forecast Re-map (MFR) algorithm, a dynamic server consolidation, was proposed by Bobroff, Kochut and Beaty (2007) that reduced the SLA violation and the capacity demand for servers. This in turn reduced the cost of running the CDC. The steps of this method consist of:

- Determining the historical data.
- Predicting the future demand.
- Re-mapping of the VM to PM.

Compared to static allocation approach, MFR reduced the SLA violation by 20%.

A mathematical model was formed to optimize the server consolidation established around LP-relaxation-based heuristic and analysing the historical workload. They extended the decision models of Bichler, Setzer and Speitkamp (2006) by applying several constraints to reduce the servers' operational costs. Optimum VM placement was realised using optimization model along with a data pre-processing approach (Speitkamp & Bichler, 2010).

*F. Greedy Algorithm*

This algorithm involves computing a locally optimum solution at each step and a globally optimum solution when the iterations are concluded. The best solution may be given at the local level, but the best solution may not be given at the global level.

According to a study by Fan, Weber and Barroso (2007) energy consumption of a server has a linear relationship with CPU usage. With this assumption, Jing, Ali and She (2103) proposed a model for the reduction of power consumption and the reduction of the number of VM placement changes. Their proposed algorithm can calculate the maximum number of required servers. Server selections are done heuristically and according to bin-packing problems. There is an upper limit to the number of bins required. A set of servers are selected for resource allocation. If the required number of servers is less

than the current number of servers, then the unnecessary one is turned off. Once the selection is made, the algorithm iterates and returns when all the demands are satisfied.
The model was tested with two sets of data – in the first group, the demand for resources is approximately steady and in the second group, the resource demand is dynamic. These two sets of data were also tested using the FFD heuristic used in the works of Tang *et al.* (2007). In this method, the resources are kept centralized in a pool. FFD heuristic dispatches the resources according to the demand to servers. Their proposed method showed slightly reduced energy consumption for both sets of data. The VM placement change was much higher with the heuristic method.
The problem, however, with this model was that the computational time increased manifold for a moderate increase in the number of services. This is given in Table 2.

**Table 2:** Number of services and the corresponding computational time

| Services | Computational Time/Sec |
|---|---|
| 100 | 70 |
| 150 | 238.77 |
| 200 | 598 |

Most of the power saving algorithms that had been developed focused on the servers but did not take into consideration the networking which consumes about 10% - 20% of the power expended by the data centre. Fang *et al*. (2013) proposed VMPlanner, which is a stepwise optimization approach, to optimize VM placement and flow of traffic by turning off as many unnecessary networks as possible. They used a Greedy Bin-Packing algorithm VMPlanner outperformed similar modules by 0.66% - 12.53%.

*G. Other Miscellaneous Algorithms*

Hosts are classified into overload, underload, and normal. Under standard circumstances, the hosts in a data centre operate at about 10% - 50% of their peak capacity. Augmenting the resource utilization of the hosts is thus necessary to improve energy consumption. Host utilization aware (HUA) algorithm was proposed by Patel and Patel (2020) to detect an under loaded host, take away the VMs, and place them in other hosts dynamically. VM selection policies and overload host detection methods used by them are given in Table 3. Their proposed model was able to shut off a number of hosts without conceding the workload requisite.

**Table 3:** VM selection and host detection policies to detect overloaded host

| **VM Selection Policies** |
|---|
| Maximum utilisation (MU) |
| Maximum correlation (MC) |
| Maximum migration time (MMT) |
| Random selection (RS) |
| |
| **Overload Host Detection Policies** |
| Median absolute deviation (MAD) |
| Interquartile range (IQR) |
| Local regression (LR) |
| Robust local regression (RLR) |
| Static Threshold (ST) |

A VM consolidation algorithm called AVVMC was proposed by Ferdaus *et al.* (2014) for stable utilization of heterogeneous resources. VM consolidation was modelled as a multi-dimensional vector packing problem (mDVPP), which is a combinatorial optimization problem where a number of VMs are packed into some number of bins or PMs. The aim was to maximize the resource utilization of active PMs and minimise power consumption. The resource utilization capturing method was based on ACO. The performance of AVVMC was compared against four other VM consolidation methods across various performance metrics and was found to lessen the resource wastage.

**Discussion:**

There are a number of methodologies that have been proposed in the past decade to address the problems of VM placement and consolidation. Existing methodologies were surveyed and a summary of the different methods, the problems solved by them along with their strengths and weaknesses, is listed in Table 4.

**Table 4:** Summary of the different VM placement algorithms along with their strengths and weaknesses

| Research Work | Proficiency | Strength | Weakness |
|---|---|---|---|
| Khosravi, Garg & Buyya, 2013 | Reduced carbon footprint by 10% - 45% and energy consumption in CDCs by 8% - 20%. | Reduction in carbon-dioxide emission and energy consumption, without compromising the QoS. | SLA violation is high. |
| Beloglazov & Buyya (2012) | Energy consumption is reduced for different levels of SLA. | Scalable, flexible, and fault-tolerant. | Not known |
| Moges & Abebe (2019) | 67% improvement in energy reduction, 78% reduction in SLA violation, 46% reduction in VM migration. | Improved energy-efficiency compared to other baseline algorithms. | Peak power of the host is unknown. |
| Feller, Rilling & Morin (2011) | Conservation of 4.7% hosts and 4.1% of energy. | Scalable and fault tolerant | Execution is slow. |
| Liu *et al*. (2014) | Reduction in the number of both PMs and VMs. | Active consumption when the number of VMs is large. | CPU and memory were the only resource under consideration. |
| Liu *et al*., 2017 | Reduction in energy consumption by 47.53%. | Scalable | CPU and memory were the only resource under consideration. |
| Dupont *et al.*, 2012 | Energy as well as carbon-dioxide emission is reduced by 18%. | Scalable and parallel computation possible. | Each host can be assigned only one VM. |
| Dupont *et al*., 2015 | Energy consumption and emission are reduced by 33% and 34% respectively. | Scalable without impacting the SLA. | Not suitable for renewable energy source. |

| | | | |
|---|---|---|---|
| Dong, Wang & Cheng (2015) | Minimization of number of active PMs, and migration cost. Dynamic VM migration. | Improved utilization of PM and network elements. | Round-trip-time and loss-rate were not taken into account. |
| Zhang, Zhuang & Zhu, 2013 | Reduction in the number of active PMs. | Improved scalability. | Impact on SLA was not taken into account. |
| Wang *et al*., 2013 | Energy consumption is reduced by 13% - 23%. | Effective for heterogeneous and large virtualized data centres. | CPU and memory were the only resource under consideration. |
| Zhao *et al*., 2014 | Reduction in power consumption, and failure rate of VM migration. | Efficient for both light and heavy load. | Connectivity of the physical nodes were not considered. |
| Kansal & Chana, 2016 | Energy saving of 72.34% for VM migration and 34.36% for the hosts. | Higher scalability and lower number of hosts. | Effective for continues<br>Optimization only. |
| Tang *et al*., 2007 | Placement changes are done one by one in isolated manner. | Execution time, demand satisfaction, and placement changes were better than any previous algorithms. | Not mentioned. |
| Mosa & Paton, 2016 | Energy consumption was reduced by 6% and SLA violation by 38%. | Works efficiently for both lightly and heavily loaded CDC. | Not tested yet on a real environment. |
| Mi *et al*., 2010 | CPU utilization and power consumption reduced by 35% and 25% respectively, compared to other. | GA-based approach can steadily converge in time and is highly scalable | CPU is the only resource taken into consideration. |
| Jing, Ali & She, 2103 | Slightly reduces the energy consumption but reduces the number of VM placement change by large margin. | Not known. | Execution is slow. |
| Ferdaus *et al*., 2014 | Resource utilization was estimated using vector-algebra. | Enhanced overall resource utilization. | Too much migration and reconfiguration. |
| Bobroff, Kochut & Beaty, 2007 | MFR algorithm that dynamically remaps VMs to PMs to optimize the number of PMs needed for a particular workload at a pre-defined SLA violation. | Adjusts to the changes in demand to migrate VMs to PMs and in the process ensures probabilistic SLA guarantees. | Relationship between multiple resources is not considered. |

| | | | |
|---|---|---|---|
| Speitkamp & Bichler, 2010 | A decision models to optimally allocate source servers or VMs to PMs while taking into account the constraints. | 31% server saving compared to other optimal allocations. | Heterogeneous resource demands for different services had a negative impact. |
| Patel & Patel, 2020 | Host Utilization Aware (HUA) algorithm was used for under-loaded host detection. | HUA efficiently detected under-loaded hosts and the freeing of hosts resulted in energy saving. Reducing SLA violation. | Not mentioned |

The state-of-the-art VM placement techniques discussed so far are mainly dynamic in nature and are based on adaptive algorithms, i.e., they can change themselves in accordance with the changing workload and demands. These methods have taken the energy consumption of the CDC into account and optimised it through strategic VM placement and consolidation. Reduction in energy consumption was most prominent with PSO, ACO, and Best-Fit algorithm-based methods, while the Greedy algorithm-based methods were the least successful. Another efficient way to reduce energy consumption was by reducing the number of active PMs. Reduction of network traffic also contributes to the amount of energy consumed by CDC. Network traffic and energy consumption are also collectively optimized in multi-objective research.

When the authors analyze the VM placement algorithms collectively, they note that they are extremely efficient in reducing the cost. However, strict server consolidation results in migration overhead. Power-based consolidation tries to use the maximum number of available resources, which leads to the compromise of QoS and SLA violation. Some of the techniques discussed here consider CPU as the primary resource while the other important resources are kept on the side-line.

**Conclusion:**

Since more than a decade, research work has been in progress for server consolidation. Most of the techniques detail the VM placement algorithms. Prime goal of these methods is to reduce the energy consumption and carbon footprint without violating the SLA. However, achieving this goal may jeopardise the QoS. Though many state-of-the-art research works have been discussed, it is not possible to rank them because each of them suggested a VM placement technique, keeping an explicit goal in mind. These techniques provided partial solutions to the problems, when studied in-depth, some trade-offs have been noted. Hence, further research work is necessary to smooth out these issues. Since the workload and the resource demands vary, the VM placement algorithms also need to be continuously upgraded. Some of the approaches the authors suggest are:

- Optimization of the trade-offs between power consumption and QoS using techniques that combine server consolidation along with load balancing.
- Implementation of the ‘Green Computing’ has been ensured by maximum resource utilization using the methods discussed above. Further research scope is there to avoid overload while fulfilling all the energy-saving criteria.
- Population-based meta-heuristic algorithms that takes into account heterogeneous resources.

**Acknowledgments:**

The authors would like to thank the administration of Brainware University, India.

**Conflicts of Interest:**

The authors declare that the research review was conducted in the absence of any commercial or economic associations that could be construed as a potential conflict of interest.

**References:**

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 Centres. *Concurrency and Computation: Practice and Experience*, *24*(13), 1397-1420.

Bichler, M., Setzer, T., & Speitkamp, B. (2006). Capacity planning for virtualized servers. In *Workshop on Information Technologies and Systems (WITS), Milwaukee, Wisconsin, USA*.

Bobroff, N., Kochut, A., & Beaty, K. (2007, May). Dynamic placement of virtual machines for managing sla violations. In *2007 10th IFIP/IEEE International Symposium on Integrated Network Management* (pp. 119-128). IEEE.

Buyya, R., Beloglazov, A., & Abawajy, J. (2010). Energy-efficient management of data centre resources for cloud computing: A vision, architectural elements, and open challenges. *arXiv preprint arXiv:1006.0308*.

Chowdhury, M. R., Mahmud, M. R., & Rahman, R. M. (2015, June). Study and performance analysis of various VM placement strategies. In *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (pp. 1-6). IEEE.

Dong, J., Wang, H., & Cheng, S. (2015). Energy-performance tradeoffs in IaaS cloud with virtual machine scheduling. *China communications*, *12*(2), 155-166.

Dupont, C., Hermenier, F., Schulze, T., Basmadjian, R., Somov, A., & Giuliani, G. (2015). Plug4green: A flexible energy-aware vm manager to fit data centre particularities. *Ad Hoc Networks*, *25*, 505-519.

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 *2012 Third International Conference on Future Systems: Where Energy, Computing and Communication Meet (e-Energy)* (pp. 1-10). IEEE.

Fan, X., Weber, W. D., & Barroso, L. A. (2007). Power provisioning for a warehouse-sized computer. *ACM SIGARCH Computer Architecture News*, *35*(2), 13-23.

Fang, W., Liang, X., Li, S., Chiaraviglio, L., & Xiong, N. (2013). VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in Cloud Data Centres . *Computer Networks*, *57*(1), 179-196.

Feller, E., Rilling, L., & Morin, C. (2011, September). Energy-aware ant colony based workload placement in clouds. In *2011 IEEE/ACM 12th International Conference on Grid Computing* (pp. 26-33). IEEE.

Ferdaus, M. H., Murshed, M., Calheiros, R. N., & Buyya, R. (2014, August). Virtual machine consolidation in Cloud Data Centres using ACO metaheuristic. In *European conference on parallel processing* (pp. 306-317). Springer, Cham.

Jing, S. Y., Ali, S., & She, K. (2013). Minimization of VM Placement Change in Energy-Aware Resource Provisioning for Cloud Data Centre. In *Applied Mechanics and Materials* (Vol. 325, pp. 1730-1733). Trans Tech Publications Ltd.

Kansal, N. J., & Chana, I. (2016). Energy-aware virtual machine migration for cloud computing-a firefly optimization approach. *Journal of Grid Computing*, *14*(2), 327-345.

Khosravi, A., Garg, S. K., & Buyya, R. (2013, August). Energy and carbon-efficient placement of virtual machines in distributed Cloud Data Centres . In *European Conference on Parallel Processing* (pp. 317-328). Springer, Berlin, Heidelberg.

Li, K., Zheng, H., & Wu, J. (2013, November). Migration-based virtual machine placement in cloud systems. In *2013 IEEE 2nd International Conference on Cloud Networking (CloudNet)* (pp. 83-90). IEEE.

Liu, X. F., Zhan, Z. H., Du, K. J., & Chen, W. N. (2014, July). Energy aware virtual machine placement scheduling in cloud computing based on ant colony optimization approach. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* (pp. 41-48).

Liu, X., Gu, H., Zhang, H., Liu, F., Chen, Y., & Yu, X. (2017). Energy-Aware on-chip virtual machine placement for cloud-supported cyber-physical systems. *Microprocessors and Microsystems*, *52*, 427-437.

Mi, H., Wang, H., Yin, G., Zhou, Y., Shi, D., & Yuan, L. (2010, July). Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centres . In *2010 IEEE International Conference on Services Computing* (pp. 514-521). IEEE.

Mills, K., Filliben, J., & Dabrowski, C. (2011, November). Comparing vm-placement algorithms for on-demand clouds. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science* (pp. 91-98). IEEE.

Mishra, M., & Sahoo, A. (2011, July). On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach. In *2011 IEEE 4th International Conference on Cloud Computing* (pp. 275-282). IEEE.

Moges, F. F., & Abebe, S. L. (2019). Energy-aware VM placement algorithms for the OpenStack Neat consolidation framework. *Journal of Cloud Computing*, *8*(1), 1-14.

Mosa, A., & Paton, N. W. (2016). Optimizing virtual machine placement for energy and SLA in clouds using utility functions. *Journal of Cloud Computing*, *5*(1), 1-17.

Patel, N., & Patel, H. (2020). Energy efficient strategy for placement of virtual machines selected from underloaded servers in compute Cloud. *Journal of King Saud University-Computer and Information Sciences*, *32*(6), 700-708.

Speitkamp, B., & Bichler, M. (2010). A mathematical programming approach for server consolidation problems in virtualized data centres . *IEEE Transactions on Services Computing*, *3*(4), 266-278.

Tang, C., Steinder, M., Spreitzer, M., & Pacifici, G. (2007, May). A scalable application placement controller for enterprise data centres . In *Proceedings of the 16th international conference on World Wide Web* (pp. 331-340).

Usmani, Z., & Singh, S. (2016). A survey of virtual machine placement techniques in a cloud data centre. *Procedia Computer Science*, *78*, 491-498.

Wang, S., Liu, Z., Zheng, Z., Sun, Q., & Yang, F. (2013, December). Particle swarm optimization for energy-aware virtual machine placement optimization in virtualized data centres . In *2013 International Conference on Parallel and Distributed Systems* (pp. 102-109). IEEE.

Wang, X., Chen, X., Yuen, C., Wu, W., & Wang, W. (2014, December). To migrate or to wait: Delay-cost tradeoff for Cloud Data Centres . In *2014 IEEE Global Communications Conference* (pp. 2314-2319). IEEE.

Zhang, L., Zhuang, Y., & Zhu, W. (2013). Constraint programming based virtual cloud resources allocation model. *International Journal of Hybrid Information Technology*, *6*(6), 333-344.

Zhao, J., Hu, L., Ding, Y., Xu, G., & Hu, M. (2014). A heuristic placement selection of live virtual machine migration for energy-saving in cloud computing environment. *PloS one*, *9*(9), e108275.