Bin packing algorithms for virtual machine placement in cloud computing: a review

Kumaraswamy S[1] **and Mydhili K Nair**[2]
[1]Department of Computer Science and Engineering, Global Academy of Technology, Bengaluru 560098, India
[2]Department of Information Science and Engineering, Ramaiah Institute of Technology, Bengaluru 560054, India

Article Info

Article history:
Received Dec 31, 2017
Revised Jul 25, 2018
Accepted Jul 29, 2018

Keywords:
Cloud computing
Virtual machine placement
Bin packing

ABSTRACT

Cloud computing has become more commercial and familiar. The Cloud data centers have huge challenges to maintain QoS and keep the Cloud performance high. The placing of virtual machines among physical machines in Cloud is significant in optimizing Cloud performance. Bin packing based algorithms are most used concept to achieve virtual machine placement (VMP). This paper presents a rigorous survey and comparisons of the bin packing based VMP methods for the Cloud computing environment. Various methods are discussed and the VM placement factors in each methods are analyzed to understand the advantages and drawbacks of each method. The scope of future research and studies are also highlighted.

Corresponding Author:
Kumaraswamy S,
Dept. of Computer Science and Engineering,
Global Academy of Technology,
Rajarajeshwari Nagar, Bengaluru - 560098 India.
+919886363412
Email: skswamy99@gmail.com

1. INTRODUCTION

1.1. Back Ground

In recent years Cloud computing architectures have received an increasing attention due to their great promises in enabling distributed computing paradigm [1]. It provides pool of computing resources enabling the Cloud user's application to use it [2]. These resources can be rented to users or customers by Cloud Service Providers (CSPs) on pay-as-you-go model, like public utility such as gas, water and electricity [3, 4].

Virtualization is an enabled technology for cloud computing. It provides the flexibility to cloud. It minimizes the energy cost, maximizes the CPU utilization, and maximizes the usage of memory/disk.

Multiple VMs on a PM may degrades the performance of PM or overutilize the PM. To overcome these issues, virtualization layer provides the mobility to place or move VM to other PMs within cloud. The plan for placing VMs among PMs plays an important role in optimizing cloud performance is called Virtual Machine Placement (VMP) which finds the optimal VM placement/mapping with PMs with various objectives and constraints.

In principle, the VMP problem is related to the classical Bin Packing (BP) problem, where the aim is to pack or place or map set of items / objects of different size into a minimum number of unit-capacity bins. It is proved in literature that both BP and VMP are NP-hard problems [5]. If there are 'm' PMs and 'n' VMs, then, m^n placement solutions. If m and n are very big numbers, solutions are huge. One can easily say that VMP is as hard as BP. The differences between Classical BP and BP-based VMP are detailed in Table 1

Since VMP problem is an NP hard problem, exact algorithms takes more time to get solutions, hence this problem can be effectively approximated using approximation algorithms such as greedy algorithms, genetic algorithms to accomplish results. These provide a solution, which though very good in most cases, may not be an optimal solution but near-optimal.

Packing algorithms like First Fit, Best Fit, Worst Fit, First Fit Decreasing (FFD), Best Fit Decreasing (BFD), Worst Fit Decreasing (WFD), greedy algorithms etc., are used to place VMs optimally. Also, authors in Ref. [6, 7, 8, 9] present few heuristics like First-Fit (FF), Best-Fit (BF), and Worst-Fit (WF) for load balancing, some of which could be tuned to serve the purpose of VMP.

In the FF, the VM is placed in the PM which is selected as first machine with available capacity. In BF, the name itself says fit to be best when left over space is least after placing the VM in the PM. In the WF, left over space is more compared to previous methods when the VM is packed in the PM. In First Fit Decreasing (FFD), Best Fit Decreasing (BFD), and Worst Fit Decreasing (WFD) heuristics VMs are sorted in decreasing order before packing. Once sorted, the VMs are placed according to original heuristics.

Table 1. Differences between Classical BP and BP-based VMP schemes

Classical BP	**BP-based VMP**
Item size is fixed once placed	Item size will vary even after placement due to SLAs
mono-objective optimization e.g., to minimize number of bins	multi-objective optimization e.g., minimize number of bins and maximize VMs performance
single-dimensional resource packing e.g., CPU or memory	multi-dimensional resource packing e.g., CPU and memory

The VMP problem can also be considered as a multi-dimensional or multi-capacity problem within a PM [10]. In Multi-Capacity Bin Packing (MCBP) problem, once item is allocated to bin, its resources like CPU, memory, bandwidth etc., cannot be allocated or available to other items in a bin.

In multi-dimensional problem, capacity is shared and is not dedicated to single item. This leads to conflict among items to get dedicated resources. Figure 1 shows the comparison between multi-dimensional and multi-capacity bin packing.

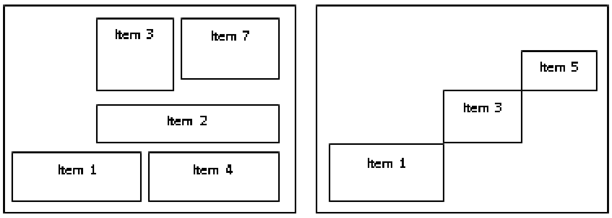

Figure 1. multi-dimensional vs. multi-capacity bin packing [10]

1.2. Problem Description

There has been considerable contributions in the literature regarding BP problem and its application for the VMP issue. Many approximation techniques to solve BP problem for VMP issue have been presented; however, comparative study on all these techniques is still lacking in the literature. Comparing all the major techniques for VMP issue through BP problem modeling, would aid in understanding the relative merits and design limitations of all these techniques, and would also aid in framing future problems in this area. The main focus of this paper is to provide comprehensive survey and simulation study on various approximation technique–regarding BP solution techniques–used for VMP issue; also, to provide the merits and limitations for these techniques, and to identify the future research problems.

1.3. Contributions

In this survey paper, the important considerations and challenges such as: resource utilization, reliability e.t.c, to efficiently design BP based solution for VMP issue are outlined. The three important approximation solutions for BP problem: First Fit, Best Fit and Best Fit Decreasing approaches, are described; also, the recently extended techniques for these approximation solutions presented in the literature are also described. Heuristics for VMP issue, usually do not guarantee performance bounds; nevertheless, they provide empirically demonstrated performance merits, and different heuristic techniques used for VMP issue are also outlined. Relative simulation oriented performance comparison for different approximation solutions for BP problem based VMP issue is presented, and corresponding merits and limitations are outlined. Finally, the future directions for the VMP issue are highlighted.

This paper is organized as follows; we start with a discussion on the considerations and challenges for VMP (Section 2), followed by BP-based problem definition in Section 3. Section 4 presents BP based VMP. Challenges of BP algorithms are presented in section 5. Section 6 presents performance evaluation of VMP schemes. This is followed by future studies (Section 7), and conclusion (Section 8).

2. CONSIDERATIONS AND CHALLENGES OF VMP

Several placement problem exist with regard to determining as to where data needs to be stored and where the job in VMs can be executed. Here we discuss the considerations and challenges of VMP.

2.1. Considerations of VMP

There are several factors involved in deciding as to when and where to place or reallocate VMs for performing computations in Cloud computing environment. The main factors are as follows.

(a) Performance: To see that the utilization of physical infrastructures are improved, data centers (DC) need to employ virtualization that could facilitate large number of applications to run simultaneously [11].

(b) Cost: Recent trend in Cloud market shows that dynamic pricing schemes utilization is being increased [12] to reduce the investment in the DC. Therefore, internal cost for VMP also needs to be considered.

(c) Locality: If accessibility and usability for users need to be considered, VMs should be closely located to users.

(d) Reliability and continuous availability: Reliability and availability of the different DC, and its expected usage frequency must be taken into account.

2.2. Challenges of VMP

Variation of scenarios in which the applications are to be deployed need relevant parameters considerations that results in the following challenges while placing VMs.

(a) Firstly, non-existence of generic model for representing various scenarios for resource scheduling.

(b) Secondly, parameterization of model for bigger problem size.

(c) Thirdly, the VMP problem is typically mapped to multiple-knapsack problem, belonging to a category of NP hard problems [13]. Thus, tradeoffs between execution time and quality of solution is an important issue to be tackled, given the size of real life DC.

3. BIN PACKING PROBLEM

Let, $S = S_1, S_2, ...S_m$ be the set of bins and all of them have the same size V. Let, $a_1, a_2, ...a_n$ be the set of n items that need to be packed in the bins. The task is to discover integer number of bins B and a B partition of the set $(1, 2, ..n)$ given by, $S_1 \cup S_2 \cup, ... \cup S_B$, so that, $\sum_{i \in S_k} a_i \leq V \; \forall k = 1, 2, ..B$. The optimal solution is to find minimal B. The integer Liner Programming formulation of Bin Packing problem is shown in Equation 1.

$$\min B = \sum_{i=1}^{n} y_i \tag{1}$$

subject to $B \geq 1$,
$\sum_{j=1}^{n} a_j x_{ij} \leq V y_i, \forall i \in (1, 2, ..n)$,
$\sum_{i=1}^{n} x_{ij} = 1, \forall j \in (1, 2, ..n)$,
$y_i \in (0, 1), \forall i \in (1, 2...n)$,
$x_{ij} \in (0, 1), \forall i \in (1, 2, ...n), \forall j \in (1, 2, ...n)$
Here, $y_i = 0$ if bin i is not used, otherwise, $y_i = 1$ and $x_{ij} = 0$ if item j is put into bin i, otherwise, $x_{ij} = 1$.

4. BP-BASED VMP

We view the problem of placement of VMs in data centers(DC)s as similar to a BP problem. We visulaize the classic BP Problem [5] in the context of VMP in the following way: VMs (objects) of different sizes are placed in hosts (bins) to ensure minimum hosts being employed for placing all VMs. In other words, the problem statement can be stated as follows: With 'x' PMs being made available with resource capacities in terms of memory, CPU and network bandwidth resources, we require 'y' VMs to be placed such that the total resource requirement of the VMs placed on a PM should not exceed its capacity. Several BP-based VMP algorithms are used in finding optimal VMP are discussed next.

4.1. First Fit (FF) approach

The First Fit Bin Packing algorithm is described in Algorithm 1. The items are scanned in any order and every item is attempted to be placed in the available bins sequentially. If it does not get fit into existing bins then, new bin is created to place the item. This algorithm achieves an approximation ratio of 2.

Algorithm 1 First Fit Algorithm

```
for $i = 1$ to $n$ do
    for $j = 1$ to $m$ do
        if Item $i$ with size $a_i$ fits in Bin $S_j$ then
            Place the item $a_i$ in Bin $S_j$.
            $j++$
        end if
        if Item $i$ does not fit in Bin $S_j$ then
            $j++$
        end if
    end for
    if Item $i$ does not fit in any Bin then
        Open a new bin and place the item in it.
        $m = m + 1$
    end if
end for
```

Authors in [14] introduce a dynamic server and consolidation management algorithm that measures historical data, forecasts future demand and re-maps VMs to PMs. To forecast future demand, time series method is used. The BP heuristic method based on FF approximation is used to perform mapping between VMs and PMs.

Authors in [15] investigated VMP problem in a Cloud as a generalized assignment problem (GAP) and formulated a multi-level generalized assignment problem (MGAP). It uses FF heuristic to solve MGAP and maximizes the profit under the SLA.

In the FF approach, the VMs are placed in the host where they first fit, according to a predefined order between active hosts. The FF algorithm [16, 5] accomplishes this by activating a single host at a time, as it is filled up with VMs. This results in DC saving energy, since only hosts that contain some VMs need to be switched on. However, since host resource usage levels tend to be close to the maximum, VM migration is more likely to happen in the presence of elasticity, that degrades the performance, besides consuming some extra energy.

4.2. Best Fit (BF) approach

The Best Fit technique is described in Algorithm 2. The algorithm works similar to First Fit, but, the items are placed in that bin which has the lowest residual capacity. The Best Fit algorithm achieves an approximation ratio of 2.

Algorithm 2 Best Fit Algorithm

```
Let $r_1 = V - 0, r_2 = V - 0, ..r_m = V - 0$ be the initial residual capacity of each bin.
for $i = 1$ to $n$ do
    for $j = 1$ to $m$ do
        if Item $i$ with size $a_i$ fits in Bin $S_j$ then
            Calculate $score_{ij} = r_j - a_i$
            $j + +$
        end if
    end for
    Fit item $i$ into that bin which has the lowest $score_{ij}$.
    $r_j = r_j - a_i$
    if Item $i$ does not fit in any Bin then
        Open a new bin and place the item in it.
        $m = m + 1$
    end if
end for
```

Authors in Ref. [17] consider greedy algorithm with two-stage to solve VMP for maximizing energy-efficiency and network performance. In first stage, if no congestion, energy optimization took priority. Authors combined minimum cut hierarchical clustering with Best Fit(BF) algorithm, enabled VMs with large traffic to be placed on the same PM or the same access switch. In second stage, if there is congestion, they applied local search algorithm to minimize Maximum Link Utilization(MLU) and link congestion with equal distribution of network traffic.

Authors in [18] consider Fat-tree DC topology to propose a solution for VMP problem and migration. The solution reduces the power cost and job delay by consolidating VMs to a few PMs and migrating VMs to close locations. In VMP stage, it considers two mechanisms, namely random placement mechanism, and power-efficient placement. In random placement, PM is selected randomly to place a VM. In power-efficient mechanism, three algorithms, i.e., FF, BF and WF are discussed; an OpenFlow controller uses these algorithms to find proper PM to deploy a VM based on the weighted difference.

The authors in [19] discuss a static disk threshold-based migration algorithm aiming at optimizing the performance of the VMs. To prevent the degradation and congestion problems in the network, authors in [20] investigate a VM placement method, called energy efficiency and quality of service aware VM placement (EQVMP). To predict the future behavior of a VM on destination host, authors in Ref. [21, 17] present a VM placement scheme named Backward Speculative Placement (BSP), that monitors the historical demand traces of the deployed VMs.

4.3. First Fit Decreasing (FFD) and Best Fit Decreasing(BFD) approach

The Fit Fit Decreasing and Best Fit decreasing techniques work similar to First Fit and Best Fit techniques respectively. But, the items are arranged in decreasing order of their sizes to improve the approximation ratio. The approximation ratio of both these techniques is < 2. Algorithms 3 and 4 describe these two techniques.

Algorithm 3 First Fit Decreasing Algorithm

```
Item Set $(1, 2, ..n)$ is arranged in decreasing order of their sizes.
$a_1 \geq a_2 \geq ... \geq a_n$
for $i = 1$ to $n$ do
    for $j = 1$ to $m$ do
        if Item $i$ with size $a_i$ fits in Bin $S_j$ then
            Place the item $a_i$ in Bin $S_j$.
            $j++$
        end if
        if Item $i$ does not fit in Bin $S_j$ then
            $j++$
        end if
    end for
    if Item $i$ does not fit in any Bin then
        Open a new bin and place the item in it.
        $m = m + 1$
    end if
end for
```

Algorithm 4 Best Fit Decreasing Algorithm

```
Item Set $(1, 2, ..n)$ is arranged in decreasing order of their sizes.
$a_1 \geq a_2 \geq ... \geq a_n$
Let $r_1 = V - 0, r_2 = V - 0, ..r_m = V - 0$ be the initial residual capacity of each bin.
for $i = 1$ to $n$ do
    for $j = 1$ to $m$ do
        if Item $i$ with size $a_i$ fits in Bin $S_j$ then
            Calculate $score_{ij} = r_j - a_i$
            $j++$
        end if
    end for
    Fit item $i$ into that bin $j$ which has the lowest $score_{ij}$.
    $r_j = r_j - a_i$
    if Item $i$ does not fit in any Bin then
        Open a new bin and place the item in it.
        $m = m + 1$
    end if
end for
```

In the simplified model, since only single dimension is considered (e.g., CPU usage), VMs could be ordered according to its resource requirement. In a complicated model, several dimensions are considered to establish a ranking amongst the set of VMs. The ranking is established in the following manner.

(a) The volume of a VM is calculated by multiplying its demands in all dimensions (by employing the volume method).

(b) The rank of each VM is determined with respect to a reference host which is, normally, the least occupied active host or the last one to be activated. In this approach, during rank calculation, the VM placement problem can be modeled as an instance of the 0-1, or Binary, Knapsack Problem [17].

For complicated model, directly we can not apply the FFD, generalization of FFD algorithm is essential. Conversion of multi dimension in to single dimension is done. Such FFD algorithms are called *FFD-based* algorithms.

Research in Refs. [22] considers Virtual Machine Monitor (VMMs) CPU cycles and migration overhead in designing VMP problem, to maximize the performance of applications and reduce the number of PMs. The evaluation of the problem is done in three different FFD based heuristics viz. Dot Product, Euclidean Distance and Resource Imbalance Vector method. It reduces the number of migrations by more than 84%.

Authors [23] considered deterministic resources viz., CPU, memory, power consumption and network bandwidth as a stochastic resource for formulating Multi-Dimensional Stochastic Virtual Machine Placement Problem (MSVP). Since it is a NP-hard problem, authors proposed polynomial time algorithm called Max-Min Multi-Dimensional Stochastic Bin Packing (M3SBP) to maximize the minimum utilization ratio of all the resources of a PMs in large scale data centers. This algorithm is inspired by First Fit Decreasing (FFD) and Dominant Resource First (DRF)

This paper [24] used data model for each VM, the data model accurately gives the resource usage of the VM. Based on this forecast, VM placement algorithm with power-aware BFD heuristic algorithm is designed. The simulation results of this algorithm show the effective reduction in power consumption, number of VM migrations, and prevent SLA violations.

Authors [25] proposed the energy-efficient and QoS aware VM placement (EQVMP) mechanism. This mechanism is a combination of three modules viz., hop reduction, energy saving and load balancing. The energy saving module uses VM placement technique inspired by Best-Fit Decreasing (BFD) and max-min multi-dimensional stochastic bin packing (M3SBP), so that it minimizes the number of server in the datacenter without SLA violation

In [26], the First Fit Decreasing algorithm is used to solve the VMP problem. In this, for each bin priority is provided. The highest priority bins consume too many resources or too few resources. So, the virtual machines in these bins will be subjected to placement.

4.4. Other Heuristics

Few of the authors consider different packing type for VM consolidation in their works. For e.g., in Ref. [5], authors consider consolidation problem as a *random variable packing* problem with bandwidth constraints, and solve it by approximation algorithm. Ref. [10] considers VM consolidation problem as multi-capacity stochastic bin packing problem. It proposes heuristic method in order to increase the efficiency of placement algorithm. [9] propose similar way of VM placement using bin packing, to increase resource utilization and profit of CSP. Few authors also consider the Next Fit (NF) approach, in which the VM is placed into the current PM (if VM resource demands are satisfied in current PM), otherwise new PM is started. It provides performance ratio of 2.

In the paper [5], authors suggest energy saving by minimizing the number of idle resources in a cloud computing environment. It has been achieved by far-reaching experiments, so as to quantify the performance of the suggested algorithm. The same has also been compared with the FCFSMaxUtil and Energy aware Task Consolidation (ETC) algorithm. The outcomes have shown that the suggested algorithm surpass the FCFSMaxUtil and ETC algorithm in terms of the CPU utilization and energy consumption.

According to Markov Decision Process (MDP) and based on reinforcement learning for auto-scaling resources, authors[27] propose an automatic resource provisioning system. Proposed system simulation results shows better performance when compared to similar approaches with respect to rate of Service Level Agreement (SLA) violation and stability.

The authors [28] proposed optimal technique based on four-adaptive threshold model to reduce energy consumption. Hosts are clustered into five clusters using K-Means Clustering algorithm viz., 1. Hosts with low load 2. Hosts with light load 3. Hosts with middle load 4. Hosts with heavy load and 5 hosts with heavy load. VMs are migrated between these clusters on the basis threshold values using mathematical modeling approach to reduce energy consumption.

Table 2 summarizes the already discussed heuristics in previous section by comparing their basic attributes like: the problem type, the type of heuristics used, the objective type (energy or bandwidth or QoS aware), type of placement, and finally objective function composite or single.

Table 2. Bin Packing Algorithms Surveyed

Paper	**Problem type**	**Heuristics**	**Energy aware**	**Bandwidth aware**	**QoS aware**	**Static/ Dynamic placement**	**Objective function**
[25]	Energy-efficient and QoS aware VMP problem	BFD	Yes	Yes	Yes	Dynamic	Dynamic
[24]	Cloud server consolidation problem	BFD	Yes	No	Yes	Dynamic	Composite
[18]	Power-efficient VM placement and migration problem	FF,BF,WF	Yes	Yes	No	Dynamic	Composite
[17]	multi-dimensional resource constraints packing problem	BF based	Yes	Yes	No	Dynamic	Composite
[15]	multilevel generalized assignment problem (MGAP)	FF	No	Yes	Static and Dynamic	single	no
[14]	Dynamic resource allocation problem	FF	NA	No	Yes	Dynamic	Single
[22]	NA	FFD-based	NA	No	Yes	Dynamic	Single
[23]	multi dimensional stochastic VM Placement MSVP	FFD-based	NA	Yes	Yes	Dynamic	Single
[10]	multi-capacity stochastic bin packing optimization problem	Hierarchical based	Yes	Yes	No	Dynamic	Composite
[5]	VM consolidation problem	Random Variable Packing(RVP)	No	Yes	No	Dynamic	Composite
[9]	Deterministic bin packing	Worst Fit	Yes	No	Yes	Static	NA
[26]	VM consolidation problem for power saving	extended FFD	Yes	No	No	rank-based	NA

5. CHALLENGES OF BP ALGORITHMS

Some of the important challenge of BP algorithms are discussed as follows.

(a) Interference between items: Due to the multi-dimensional packing of resources used by several VMs at same time, there is contention for resources among VMs. There may be affinity between two VMs due to which they may be required to be placed together. There can be various possible reasons for such affinity. For example, the network traffic between two communicating VMs may suggest that they be placed together so as to reduce the network overheads. There may be business constraints like the requirement of web-server and database layers to be together.

The contention may affect the performance of co-located VMs. There are several approaches that have been proposed to mitigate the effect of contention such as resource isolation among VMs and optimal mapping of VMs and PMs.

There could also be interference between two VMs which requires that they should not be placed together. This may be due to technical constraints. For example, say there is a resource (some remote data center) mirrored at two places, that the two VMs need to access. The VMs are programmed to access the nearest resource. Now, there may be a technical requirement to keep the two VMs separately, so that they do not access the same copy. There may be a disk contention: two VMs trying to access the same data on disk. Interference can also be due to business constraints like not keeping VMs of competitors together. Or, so that one of the VM remains available even if the physical host hosting the other VM fails.

(b) Item size and bin size: VM capacity cannot be always static over its lifetime due to SLAs. Dynamic size of VM can be considered in BP based algorithm. This is also true for bin/PM size. For variable-sized BP, there are many open questions. In terms of asymptotic worst-case ratio, following question arises:

i. which combination of sizes produces the smallest worst-case ratio?

ii. what can be said about the problem with at least three bin sizes?

iii. in terms of absolute worst-case ratio, what is a lower bound for off-line algorithms?

iv. how to design an optimal on-line algorithm?

(c) Partial packing: Existing research works do not discuss filling PMs/bins in optimal way, leaving large resource fragments or residue. Hence, resources are underutilized. Better BP algorithms need to be designed so to avoid resource fragments.

(d) No migration: most of BP based works doesn't supports migration in VMs. Tasks cannot be relocated once they have been allocated to any processor (i.e. no migration).

6. PERFORMANCE EVALUATION

In this section, the simulation analysis results for analyzing the relative performance of different VMP algorithms are presented. The simulation was carried out through open source simulation tool kit: CloudSim. Two performance metrics are utilized in simulation study: No of Severs and CPU utilization. The first metric indicates the number of servers on which all the VMs are running. It is clear that, less number of servers indicate that, efficient resource utilization has been achieved. The second metric indicates the total CPU utilization in the cloud server. Higher values of CPU utilization indicates that, proper resource utilization is being achieved.

The simulation analysis results w.r.t. no of servers is illustrated in Figure 2. It is clear that, FF, FFD and Max-Min algorithms achieve the maximum performance, mainly due to the theoretical performance bounds established for these algorithms. Similarly, NF and S-NF achieve poor performance mainly due to lack of effective theoretical performance bounds. Figure 3 illustrates the simulation analysis results w.r.t. CPU utilization. All the VMP techniques show similar performance, mainly due to their primary design focus of on achieving efficient CPU utilization.

From this simulation study, it can be inferred that, FF, FFD and Max-Min are the most favorable techniques for achieving efficient load balancing in cloud.

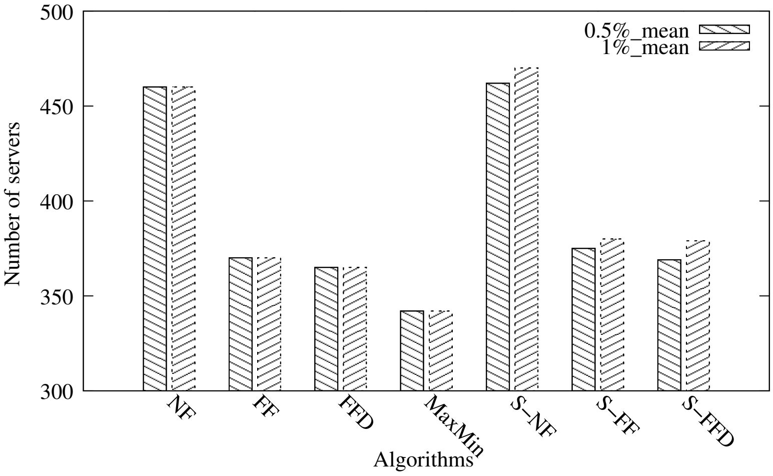

Figure 2. Number of servers vs placement algorithms

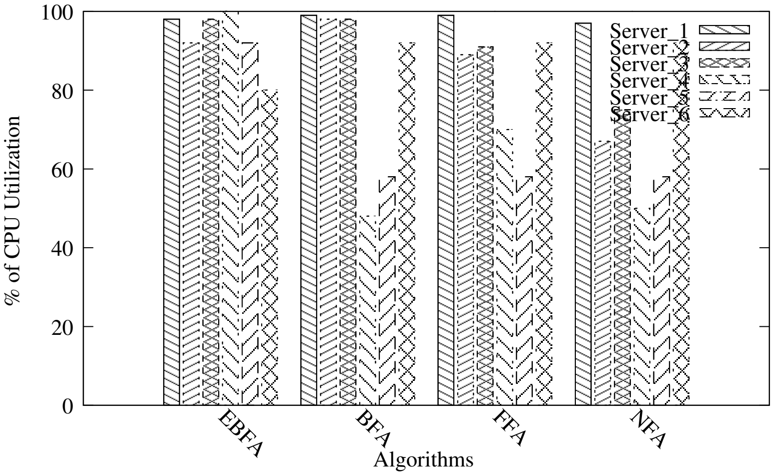

Figure 3. % of CPU utilization vs placement algorithms

7. DISCUSSIONS FOR FUTURE DIRECTIONS

This section provides VMP schemes that can be considered for the Cloud environment in the future, with respect to the following issues.

(a) Consideration of appropriate schemes.

(b) VM placement factors consideration.

(c) Resource consideration.

(d) Power, cost and network related factors.

7.1. Resource-aware VMP schemes

Managing virtual resource is an important issues in Cloud computing. Normally, for job execution, VM may require various resources. Resource-aware VMP schemes are responsible for considering the resource requirements (hardware etc.) of the VMs in the placement decisions. The optimized placement of VMs on the PMs can be achieved through efficient resource-aware placement scheme. Resource contention amongst multiple co-hosted neighboring VMs ensures maximum resource utilization in this scheme.

7.2. Power-aware VMP schemes

Due to the high cost of power consumption as well as the concerns for global warming and CO_2 emissions, the data centers need to employ Green IT for its functioning, for e.g., reduction in the number of active server, networking or other data center's components etc. The maintenance costs for the same is drastically reduced with deployment of Green data centers. To facilitate placement decisions, following factors of power-aware VM placement schemes could be considered:

(a) CPU/Server states: Four states that could be considered are idle, average, active, and over utilized [7].

(b) Network elements: Reduce network routing element's cost.

(c) Data center power usage: The cost could be divided into four categories: server base energy, server dynamic energy, cooling energy, and peak power.

7.3. Network-aware VMP schemes

The inter-data center and intra-data center network traffic can have a significant effect on revenue of the CSP owing to its dependency on the SLAs and the performance of the CSP. With the increasing trend towards communication applications, network-aware VM placement schemes could be very significant to achieve the following:

(a) Addressing the network traffic related issues with respect to VM placement.

(b) Distribution of the network traffic evenly.

7.4. Cost aware VMP

It is required to optimize the data centers maintenance cost for the CSP. Cost-aware VM placement scheme could achieve the cost saving, considering the QoS and SLAs. For better placement with cost-aware VM placement schemes, following factors need to be considered:

(a) VMs cost: The cost of switching (powering) on the VMs.

(b) PM cost: The cost of using the PM in the specific instantiation of time.

(c) Distance between the VMs and the clients: The user performance can be improved, by reducing the network distance between the VMs and the clients.

(d) Cooling cost: The cost of the cooling system in the specific instantiation in DCs.

8. CONCLUSION

In this paper, we presented a rigorous survey of the BP based VMP algorithms in the Cloud. VMP is studied and classified with reference to BP. Moreover, complete comparisons of BP inspired VMP algorithms are presented showing the items that should be considered in future research. In these comparisons, for different placement problems, BP based heuristics are used to improve energy and QoS issues in the Cloud.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, “A view of cloud computing,” *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[2] I. Foster, Y. Zhao, I. Raicu, and S. Lu, “Cloud computing and grid computing 360-degree compared,” in *Grid Computing Environments Workshop, 2008. GCE'08.* Ieee, 2008, pp. 1–10.

[3] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility,” *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[4] S. S. Manvi and G. K. Shyam, "Resource management for infrastructure as a service (iaas) in cloud computing: A survey," *Journal of Network and Computer Applications*, vol. 41, pp. 424–440, 2014.

[5] M. Wang, X. Meng, and L. Zhang, "Consolidating virtual machines with dynamic bandwidth demand in data centers," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 71–75.

[6] A. Singh, M. Korupolu, and D. Mohapatra, "Server-storage virtualization: integration and load balancing in data centers," in *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*. IEEE Press, 2008, p. 53.

[7] N. Rasouli, M. R. Meybodi, and H. Morshedlou, "Virtual machine placement in cloud systems using learning automata," in *Fuzzy Systems (IFSC), 2013 13th Iranian Conference on*. IEEE, 2013, pp. 1–5.

[8] Z. Xiao, J. Jiang, Y. Zhu, Z. Ming, S. Zhong, and S. Cai, "A solution of dynamic vms placement problem for energy consumption optimization based on evolutionary game theory," *Journal of Systems and Software*, vol. 101, pp. 260–272, 2015.

[9] K. R. Babu and P. Samuel, "Virtual machine placement for improved quality in iaas cloud," in *Advances in Computing and Communications (ICACC), 2014 Fourth International Conference on*. IEEE, 2014, pp. 190–194.

[10] I. Hwang and M. Pedram, "Hierarchical virtual machine consolidation in a cloud computing system," in *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 196–203.

[11] O. Tickoo, R. Iyer, R. Illikkal, and D. Newell, "Modeling virtual machine performance: challenges and approaches," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 3, pp. 55–60, 2010.

[12] J. L. L. Simarro, R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, "Dynamic placement of virtual machines for cost optimization in multi-cloud environments," in *High Performance Computing and Simulation (HPCS), 2011 International Conference on*. IEEE, 2011, pp. 1–7.

[13] S. Martello, "Knapsack problems: algorithms and computer implementations," *Wiley-Interscience series in discrete mathematics and optimization*, 1990.

[14] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*. IEEE, 2007, pp. 119–128.

[15] W. Shi and B. Hong, "Towards profitable virtual machine placement in the data center," in *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on*. IEEE, 2011, pp. 138–145.

[16] A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*. IEEE, 2010, pp. 577–578.

[17] J. Dong, H. Wang, X. Jin, Y. Li, P. Zhang, and S. Cheng, "Virtual machine placement for improving energy efficiency and network performance in iaas cloud," in *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 238–243.

[18] S. Fang, R. Kanagavelu, B.-S. Lee, C. Foh, and K. Aung, "Power-efficient virtual machine placement and migration in data centers," in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE, 2013, pp. 1408–1413.

[19] P. N. Sayeedkhan and S. Balaji, "Virtual machine placement based on disk i/o load in cloud," *vol*, vol. 5, pp. 5477–5479, 2014.

[20] S. Wang, Z. Liu, Z. Zheng, Q. Sun, and F. Yang, "Particle swarm optimization for energy-aware virtual machine placement optimization in virtualized data centers," in *Parallel and Distributed Systems (ICPADS), 2013 International Conference on*. IEEE, 2013, pp. 102–109.

[21] N. M. Calcavecchia, O. Biran, E. Hadad, and Y. Moatti, "Vm placement strategies for cloud scenarios," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*. IEEE, 2012, pp. 852–859.

[22] A. Anand, J. Lakshmi, and S. Nandy, "Virtual machine placement optimization supporting performance slas," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, vol. 1. IEEE, 2013, pp. 298–305.

[23] H. Jin, D. Pan, J. Xu, and N. Pissinou, "Efficient vm placement with multiple deterministic and stochastic resources in data centers," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 2505–2510.

[24] D. Dong and J. Herbert, "Energy efficient vm placement supported by data analytic service," in *Cluster,*

Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on. IEEE, 2013, pp. 648–655.

[25] S.-H. Wang, P. P.-W. Huang, C. H.-P. Wen, and L.-C. Wang, "Eqvmp: Energy-efficient and qos-aware virtual machine placement for software defined datacenter networks," in *Information Networking (ICOIN), 2014 International Conference on*. IEEE, 2014, pp. 220–225.

[26] S. Takeda and T. Takemura, "A rank-based vm consolidation method for power saving in datacenters," *IPSJ Online Transactions*, vol. 3, pp. 88–96, 2010.

[27] B. Asgari, M. G. Arani, and S. Jabbehdari, "An effiecient approach for resource auto-scaling in cloud environments," *International Journal of Electrical and Computer Engineering*, vol. 6, no. 5, p. 2415, 2016.

[28] A. Mohazabiyeh and K. Amirizadeh, "Energy-aware adaptive four thresholds technique for optimal virtual machine placement," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, 2018.

BIOGRAPHY OF AUTHORS

Kumaraswamy S is an Associate Professor in Global Academy of Technology. He is a PhD student in Ramaiah Institute of Technology, India with Master of Technology from University of Mysore (2001). He obtained Bachelor Degree in Electrical and Electronics Engineering from Bangalore University in 2009. He authored few research articles on Cloud Computing and Virtualization. His research interests are cloud computing and storage area networks.

Mydhili K Nair is working as a Professor in Ramaiah Institute of Technology, Bangalore, since 2004. She has a mixed bag of academic as well as Industrial experience both spanning close to a decade each. In the IT Industry, she has adorned various roles ranging from Technical Lead to Project Manager in different IT companies. She has won the prestigious IBM Faculty Award in the year 2011 for her collaborative research association with IBM. She has authored many book chapters and Scopus indexed research papers in reputed publishing avenues such as Springer, CRC Press, IEEE etc. She has been active in IEEE Women in Engineering forums in organizing international conferences of repute, chairing sessions and give invited talks. She and her students have won many project competitions and best paper awards at State and National level.