Journal of Network and Computer Applications 66 (2016) 106–127

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

An overview of virtual machine placement schemes in cloud computing

Mohammad Masdari, Sayyid Shahab Nabavi, Vafa Ahmadi

Department of Computer Engineering, Urmia branch, Islamic Azad University, Urmia, Iran

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form
18 December 2015
Accepted 15 January 2016
Available online 29 January 2016

Keywords:
Virtual machine
Migration
Placement
Power efficiency

ABSTRACT

In cloud computing, Virtual Machine (VM) placement is a critical operation which is conducted as part of the VM migration and aimed to find the best Physical Machine (PM) to host the VMs. It has a direct effect on the performance, resource utilization and power consumption of the data centers and can reduce the maintenance cost of the data centers for cloud providers. Numerous VM placement schemes are designed and proposed for VM placement in the cloud computing environment aimed to improve various factors affecting the data centers, the VMs and their executions. This article provides a complete survey and analyses of the existing state of the art VM placement schemes proposed in the literature for the cloud computing and data centers. Furthermore, it classifies the VM placement schemes based on the type of the placement algorithm and assesses their capabilities and objectives. Moreover, the properties, advantages and limitations of the VM placement schemes are compared. Finally, the concluding remarks and future research directions are provided.

Contents

E-mail addresses: M.Masdari@iaurmia.ac.ir (M. Masdari), Shahab.nab@gmail.com (S.S. Nabavi), Wafa_ahmadi@yahoo.com (V. Ahmadi).

http://dx.doi.org/10.1016/j.jnca.2016.01.011
1084-8045/

1. Introduction

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable virtual resources which can rapidly be provisioned and used based on the pay-for use model. It provides facilities to its users to dynamically scale up the applications, the platform and the hardware infrastructure (Gahlawat and Sharma, 2014; Tang et al., 2014; Shu et al., 2014). The cloud service models can be categorized into Software-as-a-Service (SaaS), Platform-as a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) (Buyya et al., 2009; Lombardi and Di Pietro, 2011; Zhang et al., 2010).

In addition, some primary deployment models as public clouds, community clouds, private clouds, and hybrid clouds exist (Komu et al., 2012; Davidovic et al., 2015; Srinivasan et al., 2015; Zhang et al., 2013).The Public clouds provide their services via the Internet to everyone, while private clouds are internally managed and operated by a single organization and offer increased reliability, control, and security to their users. On the other hand, a community cloud is deployed between two or more organizations with similar service requirements. Moreover, the hybrid clouds are a collection of the clouds which possibly of different types and interconnected with each other (Komu et al., 2012; Zhang et al., 2013). The cloud computing provides numerous advantages to its users and providers. Some of the key benefits of the cloud computing are flexibility, cost savings, location and hardware independence, multi-tenancy and reliability.

The cloud computing enables its users to utilize computing resources from the cloud data centers as a service, instead of owning it. For this purpose, cloud computing applies virtualization in which hardware resources of one or more computer systems are divided into several execution environments called Virtual Machines (VMs). Generally, each VM is isolated from other VMs and can act as a complete system to execute the user applications. In order to host a VM, a Physical Machine (PM) or server must provide all resources that the VM requires, including its CPU, memory, storage and network bandwidth requirements (Vu and Hwang, 2014). Inside a server or a physical machine, the VMs are controlled by a layer of software called VM Monitor (VMM) or hypervisor, which, as indicated in Fig. 1, resides between the hardware platform and the VMs. Generally, the VM monitor supports the creating, migrating and terminating VM instances (Komu et al., 2012).

VM migration is an interesting cloud computing feature which is aimed to respond to dynamic requests of the VMs in order to guarantee the promised Service Level Agreement (SLA) to the cloud consumers. Thus, when a VM requests some resources which cannot be provided in the hosted physical machine, the VM is migrated to another physical machine to satisfy the VM's requested resource. Also, the VMs may be migrated to provide better management of the physical machines and data centers.

One of the important operations which is conducted as a part of the VM migration is the VM placement in which a proper physical machine is selected to host the VM. Because of its importance, the problem of the VM placement in the cloud computing environment has been the focus of substantial researches, and numerous VM placement schemes have been designed and proposed in the literature. Each VM placement scheme pursuits some objectives and utilizes a placement algorithm which applies some placement factors to achieve these objectives. For example, some VM placement schemes try to reduce the data center's network traffic and power consumption. Also, an ideal VM placement scheme should reduce the need for future VM migrations and improve the resource utilization and availability of the data centers' resources (Dashti and Rahmani, 2015).

This paper presents a comprehensive survey and analysis of the various state of the art the VM placement solutions have been proposed for the various cloud computing environments and data centers. The proposed schemes are classified according to the algorithm type applied in the VM placement, and the properties, advantages and limitations of each scheme are investigated in detail. Also, besides analyzing the objectives of each VM placement scheme, we have investigated and compared various placement factors which have been utilized to realize each objective. Conducting this research is very important for understanding the current trends in the VM placement methods and designing effective VM placement for emerging issues of the cloud computing environment.

The rest of the paper is organized as follows: Section 2 illuminates the VM placement problem in cloud computing and Section 3 provides an overview of proposed VM placement schemes. Section 4, discusses about the simulators/provider applied in VM placement schemes and Section 5 presents an in-depth comparison of VM placement schemes. Finally Section 6 provides the concluding remarks and future research directions.

2. Problem definition

Virtualization is one of the core technologies of the cloud computing environment. A Virtual Machine (VM) emulate a computer system and presents the targeted system and support the execution of an operating system. Classification of the VMs can be based on the extent of which they implement functionality of the targeted machines. Various levels of virtualization are as follows:

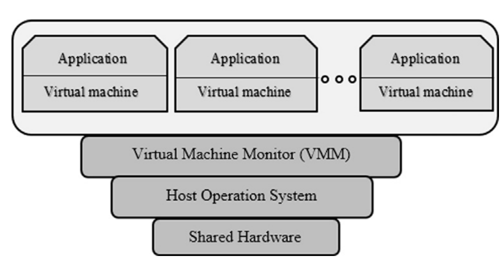

Fig. 1. Hosted virtualization.

- Full virtualization
- Para virtualization
- Hardware assisted virtualization
- Resource virtualization

Generally, each VM shares physical hardware resources with the other VMs and to be more secure, it is kept isolated from other VMs (Zhang et al., 2013). As shown in Fig. 1, the VMs are managed by a hypervisor or the VM monitor (VMM) that is a software, firmware or hardware which creates and runs the VMs. The hypervisor presents the VMs with a virtual operating platform and tries to supervise the VMs' operating system.

The VM migration is one of the main operations of virtualization in cloud computing which allows the movement of the VMs from one host to another. Migration of a VM requires the source and destination host machines to coordinate the transfer of the VM's definition and state files. The VM migration provides the ability to adapt the placement of the VMs to satisfy performance objectives, resource utilization and communication locality, reducing performance hotspots, energy consumption achieving and fault tolerance.

The VM migration process consists of four steps. First, the physical machine, over loaded or under loaded, is selected, then one or more of its VMs are selected for migration. Afterwards, the VM placement is performed for which a physical machine is selected to host the VM and satisfy its resource requirements including bandwidth, CPU utilization, memory and storage. Finally, VMs are migrated to the destination physical machine. Current virtualization technology allows Non-live migration and Live migration which the Live migration provides the ability to easily relocate a VM from one physical machine to another without shutting it down. Live migration schemes utilize the following phases:

- *Push phase*: The VM's source host's VMM pushes certain memory pages across the network to the destination physical machine while the VM is running. Consistency of the VM's execution state is ensured by resending any modified pages during this process.
- *Stop-and-copy phase*: The source host's VMM halts the running VM and copies all the memory pages to the destination physical machine. Then, the new VM is started.
- *Pull phase*: The new VM runs in the destination physical machine and when a page is accessed that has not yet been copied, a page fault occurs and it is copied from the source host.

The VM placement is a critical operation which is conducted to determine the most appropriate PM or server to host the VM. Selecting a suitable host is very important to improve power efficiency, resource utilization and QoS support in a cloud computing environment.

The VM placement in the cloud computing is very important and has been proved to be a highly complex task (Gahlawat and Sharma, 2014). Because, firstly, the arrival patterns of the VM instantiation requests may not be predictable. Secondly, since the size of the data center is often large and for a given load, finding the optimal or near optimal scenario has been shown to be NP hard. Because, for example, if n is the total number of the VMs and m is the total number of servers, then the number of possible mapping can be m^n. The problem of the VM placement can be divided into two parts: the first part is the admission of new requests for the VM provisioning and placing the accepted VMs on hosts (the initial VM placement), whereas the other part is the optimization of the VMs placement in the VM migration process (Fig. 2).

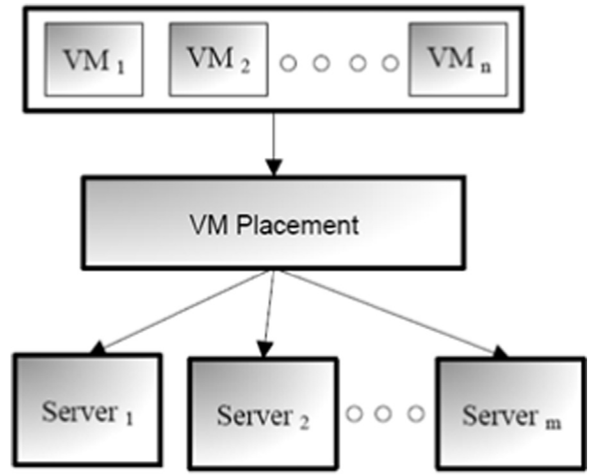

Fig. 2. VM placement.

Generally, to achieve a better performance, the VMs are migrated when some server or physical machine is under-utilized or over-utilized. A data center is composed of a number of physical machines or servers. As indicated in Fig. 3, sometimes the data center's servers may not be used to their full capacity which this issue reduces the energy efficiency of the data centers and increases their power consumption costs. For example in Buyya et al. (2010), it is specified, due to the energy consumption of the various hardware components, an idle server may consume about 70% of the energy it consumes at a full CPU speed. To reduce this problem, the VMs should be consolidated and the numbers of active servers must be decreased. This improves the resource utilization and availability of the cloud resources and reduces the power consumption of the data centers. For example, in Fig. 3, by using the VM consolidation, the number of active servers is reduced and the VMs are packed in only two remaining active servers. The other two servers can be shut down and later they may be turned on by using methods such as Wake on LAN when the demand for the resource increases. Also, to improve the performance and meet the SLAs, sometimes the VMs are migrated from an overloaded server to reduce the server load. For this purpose, first the overloaded host should be found and some VM in the overloaded host should be selected for migration and placed on some destination lightly loaded physical machine. Figure 4 indicates the classification of the VM placement schemes.

VM placement schemes can be classified as dynamic and static:

- Static VM placement: in which the mapping of the VMs is fixed throughout the lifetime of the VM and it will not recomputed for a long period of time.
- Dynamic VM placement: in which the initial placement is allowed to change due to some changes in the system load or etc.

Furthermore, the dynamic VM placement algorithms can be categorized as reactive and proactive VM placement:

- Reactive VM placement: which makes change to an initial placement after the system reaches a certain undesired state. The change may be made because of the performance, maintenance, power or load issues or some SLA violations.
- Proactive VM placement: which changes the VM's Physical Machine (PM) of an initial placement before the system reaches a certain condition.

VM placement assists in the effective management of the data center's resources and considers factors such as load balancing and

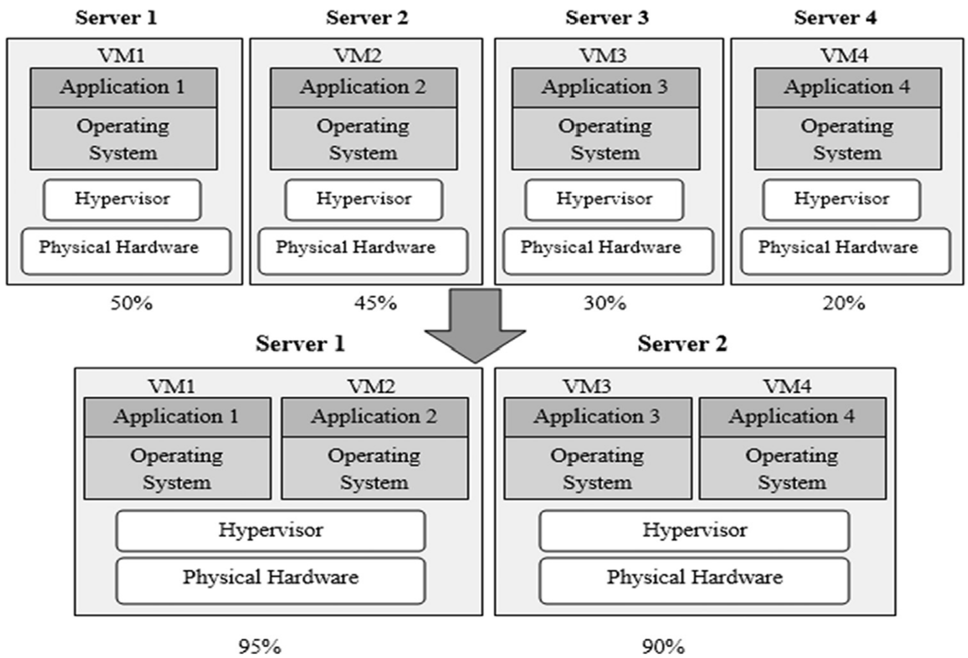

Fig. 3. An example of VM consolidation.

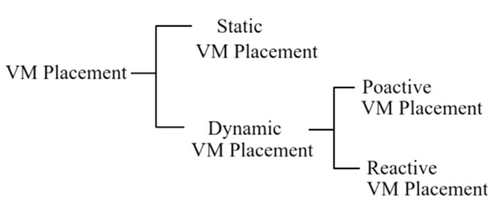

Fig. 4. Classification of VM placement schemes.

infrastructure maintenance without compromising application availability. To present more efficient VM placement and select the more appropriate server to host the VM, many factors related to the data centers, the VMs and their relation should be considered. Also, even some placement schemes use prediction algorithms which forecast the future resource demands to provide better placement decisions. Also, VM placement algorithms can be classified based on the type of cloud, as follows:

- VM placement in a single cloud
- VM placement in federated clouds which the owner of a volunteered resource may withdraw from the federation at any time.

A VM placement algorithm in volunteer cloud federations consists of two parts. First, an individual member cloud should be found to host the VM (intra-cloud VM placement) and second, it must be determined which host of the selected cloud must host the VM (intra-cloud VM placement). Thus, the VM placement in federated clouds is more complicated in volunteer cloud federations. The objectives of an ideal VM placement can be listed as follows:

- Reducing intra-data center traffic
  - Decreasing VM to VM traffic
  - Decreasing VM to storage(data) traffic
- Minimizing inter-data center traffic in federated cloud
  - Decreasing VM to VM traffic
  - Decreasing VM to storage(data) traffic
- Preventing congestion in data center's network
- Distributing traffic evenly in data center's network
- Mitigating energy consumption
- Reducing cost for cloud providers and increasing the return on investment (ROI)
- Increasing security
- Maximizing resource utilization: The effective and efficient utilization of each server reduces the need for more PMs to host VMs.
- Improving Load balancing
- High performance
- Locality: To increase accessibility, the VMs should be located close to the users, but this may not necessarily produce an optimal placement.
- High reliability and availability: To achieve these properties, the VMs may be placed, replicated or migrated across multiple geographical zones. During this procedure, factors such as the importance of the data/service encapsulated in the VMs, its expected usage frequency, and the reliability of the data centers must be considered.
- Minimizing the number of active servers: The energy consumption of a data center is mainly caused by running PMs, thus VM placement schemes try to minimize the number of servers which are in use.
- Reducing the number of active networking elements (switches)
  - Reducing the number of active ports
- Minimizing the SLA violation: SLA is usually defined as the difference between the total MIPS requested by all the VMs from the total MIPS allocated to its VMs

The problem is how to achieve most of these properties in a VM placement scheme, to have more efficient, low overhead, low cost and scalable VM placements in the cloud data centers. However, some of these objectives are in contradiction with each other and all of them may not fully be achieved in one placement scheme. As a result, as it will be indicated in next sections, each placement scheme only focuses on some of these objectives.

3. Analysis of VM placement scheme

This section discusses the existing VM placement schemes proposed in the literature for cloud IaaS layer. Almost all of the proposed VM placement schemes consider a three-tier architecture for data

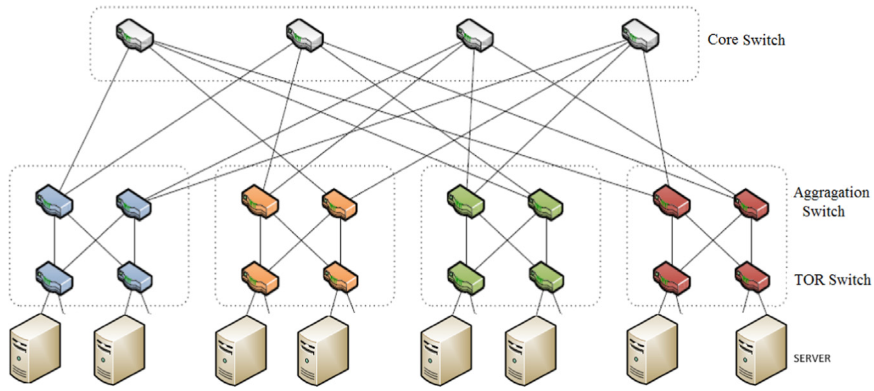

Fig. 5. Topology of data center network.

center network. The goal of data center network is to interconnect several servers with high-speed links. As indicated in Fig. 5, at the bottom level servers are organized in racks, and each server in a rack connects to one or two Top-of-Rack (ToR) switches. Also, each ToR switch connects to one or two switches at the aggregation tire. Finally, each aggregation switch connects with multiple switches at the core tier. Most new architecture designs such as Virtual Layer 2 (VL2) (Jayasinghe et al., 2011) and Fat-Tree (Zamanifar et al.) share similar richly-connected topologies and differs more on how addressing and routing are implemented.

This section provides an elaborate assessment of each placement scheme and highlights their objectives and the placement factors which have been considered in each solution. Figure 6 presents the classification of VM placement schemes proposed for cloud environment. Numerous simple and primary VM placement algorithms exist which are widely used by other placement schemes for comparison and evaluation. These algorithms are as follows (Gupta and Pateriya):

- First Fit: It is a greedy approach which scheduler considers the PMs sequentially, one by one, and places the VM to the first PM that has enough resources.
- Next Fit: This placement method considers the PMs one by one and places the VM to the second PM which has the required resources.
- Random Fit: A random physical machine is chosen for placing the VM.
- Least full first: The physical machine which is least full and satisfies the resource requirement of the VM selected.
- Most Full First: The physical machine which is most full and has the resource requirement of VM is selected.

First Fit Decreasing product: Each physical machine or the VM is represented with different resource capacities and each resource capacity is normalized in the percentage of utilization.

3.1. Constraint programming-based VM placement

Constraint programming is a programming paradigm which has been applied for VM placement in cloud computing environments. It is useful for combinatorial search problems and its solutions must satisfy the constraints on relations between variables.

In Van et al. (2010) the authors propose a constraint programming-based resource management framework combining a utility-based dynamic VM provisioning manager and a dynamic VM placement manager. In this scheme, the VM provisioning and placement problems are expressed as two constraint satisfaction problems. This scheme uses the Entropy resource manager for homogeneous clusters, which performs dynamic consolidation based on constraint programming and takes into account both the problem of allocating the VMs to the available nodes and the problem of how to migrate the VMs to these nodes. To reduce energy consumption and considering performance requirements, this scheme proposes a flexible framework for the allocation or reallocation of the VMs in a data center which computes and enacts the best possible placement of the VMs based on constraints expressed through the SLAs.

In Jayasinghe et al. (2011) the authors propose a Structural Constraint-Aware VM Placement (SCAVP) to improve the performance and availability of cloud services. This scheme proposes a hierarchical approach consisting of four approximation algorithms to solve the placement problem and supports the following constraints:

- Demand constraints: which provide required resources to the service to meet its SLA.
- Communication constraints: which decrease communication costs by assigning VMs with more communication to near servers.
- Availability constraints: permit the availability of application to be improved, by deploying the VMs across different isolation levels of the datacenter.

The placement engine is the main component of SCAVP and is composed of the Structure-Aware Planner (SAP) and the Demands-Aware Planner (DAP) where the SAP deals with the structural constraints and the DAP handles the resource constraints. Moreover, the SAP produces a candidate placement plan, which the DAP refines. SCAVP uses a divide-and-conquer approach to create the placement plan. In the divide stage, the application graph is split into a set of smaller VM groups, and the group size is specified with the capacity of a server rack. Then for each VM group, the SAP finds a physical machine rack.

In Dupont et al. (2012) the authors present a flexible framework which focuses on the energy-aware allocation/consolidation of the VMs in cloud data centers and uses constraint programming to express data center constraints. The core element of the framework is the optimizer which attempts to reduce the total communication traffic and optimize network Maximum Link Utilization (MLU). On the condition of slight variation of the total traffic, this scheme can balance network traffic distribution and reduce network congestion hotspots. The optimizer aims at computing an assignment of the VMs to the nodes that minimize the overall energy consumption of a federation of data centers while satisfying different SLAs.

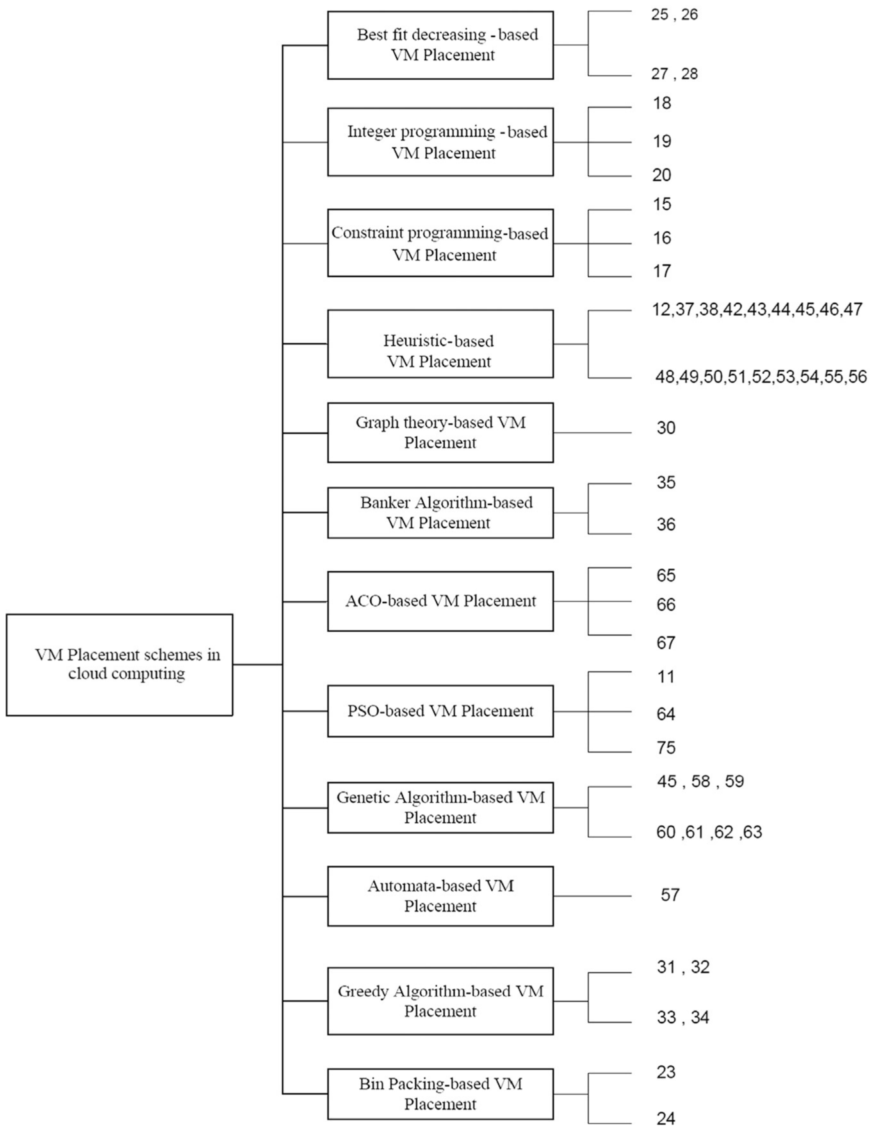

Fig. 6. Classification of VM placement schemes in cloud environment.

3.2. *Integer programming-based VM placement*

Stochastic programming is applied for modeling optimization problems that involve uncertainty. Chaisiri et al. (2009) propose OVMP or Optimal VM Placement algorithm to provision the resources of multiple cloud providers and reduce the cost for hosting the VMs considering future demand and cost. This algorithm utilizes the Stochastic Integer Programming (SIP) to rent resources from providers and the trade-off between the advance reservation of resources, and the allocation of on-demand resources is tuned to be optimal. This scheme consists of two stages which, at the first stage, defines the number of the VMs provisioned in the reservation phase, while, at the second, defines the number of the VMs allocated in both utilization and on-demand phases. In other words, the second stage represents the actual number of the VMs required by the user and actual prices defined by providers.

Generally, they proposed VM-placement schemes, based on their objectives can be classified as traffic-aware, energy-aware, cost-aware, resource-aware and application-aware schemes. Table 1 indicates the objectives of Integer programming-based and constraint programming-based VM placement schemes.

In Tordsson et al. (2012) the authors provide a cloud broker which tries to optimize the placement of virtual infrastructures according to user-specified criteria across multiple clouds. The algorithm is based on the integer programming formulation and enables price–performance placement tradeoffs. Also, users can steer the VM allocation by specifying maximum budget and minimum performance, as well as constraints with respect to load balance, and hardware configuration of the individual VMs.

In Li et al. (2012) authors present an approach to optimal VM placement within datacenters for predictable workload scenarios. They formulate the problem as Min–Max optimization and present a method for load balancing based on the integer programming. Moreover, they perform some pre-processing before solving the optimization problem. Upper bound based optimizations are used to reduce the time required to compute a final solution, enabling larger problems to be solved. For further scalability, three approximation algorithms are presented for trade-offs in scalability and performance, based on heuristics and/or greedy formulations.

In IaaS cloud environment, the VMs may require access to data files stored in the distributed storage centers across the cloud. The performance of such accesses depends on the data transfer delay, data files size, location of the VMs that run the applications and

Table 1
Objectives of Integer programming and constraint programming-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Zamanifar et al.		✓			
Tordsson et al. (2012)			✓		
Dupont et al. (2012)		✓			
Zhang et al. (2014)			✓		
Jayasinghe et al. (2011)	✓				
Chaisiri et al. (2009)			✓		
Gupta and Pateriya					

Table 2
Properties of Integer programming and constraint programming-based VM placement schemes.

Schemes	SLA Support	Security	Environment	Algorithm type
Zamanifar et al.			Multiclouds	ROVMP
Tordsson et al. (2012)			Heterogeneous	Integer programming
Zhang et al. (2014)	✓		Tree-like DC topology cloud	Multi objective nonlinear programming
Jayasinghe et al. (2011)	✓		IAAS Cloud	Constraint programming SCAVP
Chaisiri et al. (2009)			Multiple cloud	Integer programming
Gupta and Pateriya)			Private cloud	Constraint programming

allocation of data rates to the VMs. To reduce this problem, in Zamanifar et al. authors propose ROVMP or Rate Optimal VM Placement algorithm to be executed in broker that determines the optimal placement and allocates optimal data rates to VMs to minimize data transfer delay. In this scheme, the problem is converted to a linear mixed-integer programming problem and a branch-and-bound technique is employed to efficiently find the optimal solution. This scheme compares its results with the random placement algorithm and another placement scheme which applies the TCP Vegas and places the VMs one-by-one without taking the optimal rate allocation into account for the VMs.

Table 2 exhibits the properties of the Integer programming and constraint programming-based VM placement schemes. In this table, the algorithm applied for VM placement and the simulation environment is specified. Also the security and SLA support by each VM placement scheme is determined.

3.3. Multi-objective nonlinear programming-based VM placement

In Zhang et al. (2014) authors address the SLA aware cost efficient VMs placement problem in cloud. They consider the heterogeneous servers and the random requirements of the VMs and model the efficient and cost-aware VM placement as a multi-objective nonlinear programming. In this scheme, VM partition and statistic multiplex as well as similarity tactics are used to provide a novel offline, and an online heuristic algorithm is presented which can guarantee the QoS while saving. In this scheme, by exploiting topology information of the data center, VM cluster with higher traffic is made to stay together to reduce the communication delay while saving the inter-server bandwidth consumption, and the VMs may be independent or correlated. For this purpose, statistic multiplex and similarity techniques are used to consolidate the VMs and to compact more of them into one PM. By using these methods, the resource usage of the PMs is reduced and the resource fragment leak in the PMs is minimized. Also, in this scheme the violation of resource capacity is kept minimal and thus QoS is not deteriorated.

3.4. Bin packing-based VM placement

Some VM placement schemes consider the VM placement problem as a bin packing problem which is an extension of the first fit decreasing. However, as the bin packing problem is NP-hard, an ordinal method cannot be applied to solve it in an acceptable complexity, thus heuristic methods are suggested. For example, in Xiaoli and Zhanghui (2012) the authors improve online bin packing algorithm and propose a new energy-aware approach to consider energy efficiency and resource utilization in a cloud datacenter. In addition, they present an over-provision approach to deal with the varying resource demands of the applications and present an essential support for such applications to get economical running. As indicated in Fig. 7, this scheme is composed of Event invoker, Concentrate manager and resource pool components. The event invoker receives the feedback information such as VM starting, VM resizing, VM migrating and VM departure from the resource pool. The Concentrate manager receives the commands from Event invoker, and makes the VM placement scheme according to the assignment algorithm. Energy cost in the model above is caused by the VM running, VMs migrations and the fixed consumption of starting servers. The resource size and migration number determine the wastage of energy caused by migration operations. So, the smaller the resource size and the smaller the migration number, the less energy will be consumed. The energy consumption for opening a server is fixed and unmodifiable by the algorithm.

Another bin-packing-based approach for VM placement is presented by Babu and Samuel (2014) which tries to increase the datacenter's resource utilization and improve the cloud provider's profit. In the proposed work, Bin-Packing strategy based approach is used for VM creation. The cost of running a datacenter can be minimized by allocating maximum number of the VMs to particular servers. In this scheme, each PM in the datacenter is considered as bins and the VMs as the objects to be filled in the bin. This bin packing method tries to minimize the number of servers required for placing user requested VMs and also reduces the job completion time. Also, in this scheme, the cloud broker employs worst-fit method for VM placement and it places the VMs to less active servers and constrains the mapping of the VMs, such that power consumption and computation cost are minimized. In addition, for VM placement, in worst-fit algorithm, the first PMs and VMs are sorted in the decreasing order of utilization and the required MIPS. Then, the first PM is founded which has the required resources from the list of the sorted PMs to deploy the first VM from the list of sorted VMs. This procedure is repeated till all the VMs are mapped to the PMs.

Table 3 shows the objectives of the bin packing-based and best fit decreasing-based VM placement schemes. As it is indicated in this table, the energy and resource related objectives are more considered in these schemes.

3.5. Best fit decreasing-based VM placement

The authors Noumankhan Sayeedkhan and Balaji (2014) propose a disk I/O load-based VM placement algorithm called FFDL and provide a Static Disk Threshold-based migration algorithm to optimize the performance of the VMs. In FFDL, the system model consists of the global and local managers which the local managers are part of the VMM positioned on each node and continuously check when to migrate VMs. The user issues its request with some CPU parameters to the global manager which in turn announces the VMM for VM placement. The local manager sends a report to the global manager about the utilization check of its node and to check the overall resource utilization. In this scheme, the VM placement is performed based on best fit decreasing algorithm which first sorts the VMs in the decreasing order of its VM utilization and for each host in the host list, if the host has enough resource for the VM, then selects the host as a destination, otherwise it does nothing.

The authors in Wang et al. (2014) investigate an improved VM placement method, called EQVMP or Energy efficiency and Quality of Service aware VM Placement, to prevent the unexpected congestion and degradation in the network. As indicated in Fig. 8, EQVMP is a three-tier algorithm that considers hop reduction, energy saving and load balancing in VM placements. It switches the VMs on and off for the purpose of energy saving and balancing the traffic load. In the first phase of this scheme, the VMs are partitions based on the collected traffic loads and datacenter topology architecture to reduce traffic transmission among groups by graph partitioning. Then, it decides the minimum number of server without Service-Level Agreement (SLA) violation and, finally, the Open Flow controller assigns the paths to avoid congestion and balance the network load. In the second phase, the VMs are placed into servers by considering their resource demands such as CPU consumptions, memory usages and bandwidth demands. The placement is based on the Best Fit Decreasing (BFD) and Max–Min Multidimensional Stochastic Bin Packing (M3SBP). Then, energy saving process minimizes the number of power-on servers. After that, the Open Flow controller balances the traffic flow to avoid congestion based on the network status.

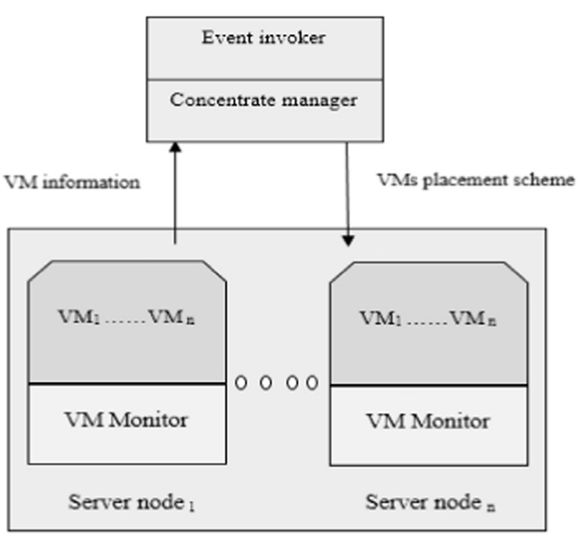

Fig. 7. Resource placement model in cloud computing.

In Calcavecchia et al. (2012), the authors present a VM placement scheme named Backward Speculative Placement (BSP) which by monitoring the historical demand traces of the deployed VMs, projects the future behavior of a VM on a target host. This solution performs the placement decisions in the continuous deployment phase and ongoing optimization phase. The algorithm applies a decreasing best fit strategy according to the "demand risk" score. Also, in ongoing optimization phase, the current placement is re-optimized by relocating the VMs to the other hosts. Actions produced during the continuous deployment phase provide a sub-optimal allocation of the resources to the VMs. To improve the placement, an optimization task consisting VM migrations is periodically performed.

Some VM placement approaches decrease the energy consumption, but ignore the issue of aggressive consolidation, which may lead to the network performance degradation. In Dong et al. (2013) the authors propose an optimal VM placement scheme which considers multiple resource constraints. When meeting the constraints of PM resources and network link capacity, cross-optimizing VMs are placed on the PMs for avoiding network congestion. It can maximize the resources utilization of the PMs and network elements to reduce SLA violation and allow the idle resources in a sleeping state to reduce energy consumption in the IaaS cloud. In this scheme, an energy model is proposed, which considers the energy consumption of physical servers, and includes CPU, memory, storage, and network interface power consumptions. Also, energy consumption of the network elements is considered in the energy model which consists of power consumption of the network elements including switches and active links. They propose a two-stage greedy algorithm which combines the minimum cut hierarchical clustering with the best fit algorithm that enables the VMs with large traffic to be placed on the same PM or the same access switch, to reduce the network traffic. Then best fit method is applied to minimize the number of the PMs, and the local search algorithm is utilized to minimize the maximum link utilization and to keep a balanced distribution of the network traffic to reduce the link congestion.

Table 4 shows the algorithm applied for VM placement, the simulation environment, the security and SLA support by each VM placement scheme in bin packing-based and best fit decreasing-based VM placement schemes. As indicated in this table, nost of these VM placement scheme are aimed to support the SLA and user requested QoS.

Table 3
Objectives of bin packing and best fit decreasing-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Xiaoli and Zhanghui (2012)		✓			
Babu and Samuel (2014)		✓		✓	
Noumankhan Sayeedkhan and Balaji (2014)				✓	
Wang et al. (2014)		✓			
Calcavecchia et al. (2012)				✓	

3.6. Game theory-based VM placement

Game theory addresses the situations in which the outcome of a person's decision depends not just on how they choose among several options, but also on the choices made by the people they are interacting with.

In Xiao et al. (2015) authors give an evolutionary game theory-based algorithm to deal with dynamic VM placement. In this scheme, first, a model of energy consumption is built to provide support for computing the amount of energy consumed during the complicated process of dynamic adjustment of the VMs placement. This scheme takes the initial solution into account and produces an executable list of VM live migrations from the initial solution to the target solution. It is based on multiplayer random evolutionary game theory which involves a population of decision makers, and some of whom may give up their best strategy, while choosing a random strategy at a certain rate. In this scheme, the VMs are mapped into multiplayers that take part in the evolutionary game and in each round, all the players choose a PM to stay, to maximize their returns which is decreasing amount of energy consumption. An equilibrium point will be reached at the end of the round. If the utility of the current solution is better than the best solution so far, the solution will be adopted for the next round of the game. To avoid obtaining partial optimal solutions, some VMs will not choose the best PM, and rather a random PM according to a stochastic perturbing probability. In this scheme, each PM can have four different states which include running, ready, sleep and off. The ready state means that PM is on but there is no VM running on it. As the PM in this state still consumes energy, keeping a PM in such a state for a long time will waste energy, and hence it should be switched to the off state. While the state transition cuts down the energy waste extra time and energy will be needed to turn it back on.

Consequently, if the switches between on and off are too frequent, it will cost more energy, and may even bring down the performance level of the system. To this end, a trade-off option, the sleep state, is introduced, on which the PM consumes very little energy close to zero, yet it can be switched to the ready state instantly with very low energy cost.

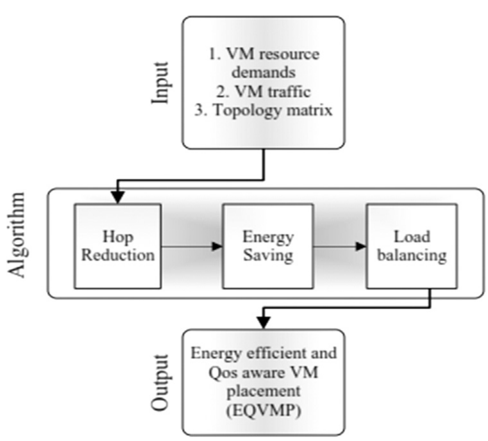

Fig. 8. EQVMP VM placement.

Table 5 indicates the objectives of Game algorithm-based, Graph theory-based and Greedy-based VM placement schemes. As shown in this table, energy is one of important factors which have been considered in these schemes to achieve more green cloud computing.

3.7. Graph theory-based VM placement

Malicious VMs could attack the other VMs, their hypervisors or hosts and may infect them. Moreover, an infected VM can migrate to other hosts and can infect other VMs too, making the infection epidemic. In Abazari and Analoui (2014) the authors present a scheme based on the combination of graph theory and agent based simulations to study epidemics in the cloud environment and control the spread of infections. They analyze the impact of VM placement on the cloud security and consider both the malicious intentions of the VM for infecting sibling machines and spreading among as many physical machines as possible. They apply the concept of "Epidemic Theory" in the VM placement and estimate the amount of the infected machines based on the proposed collaboration graphs in cloud infrastructure.

3.8. Greedy-based VM placement

Kanagavelu et al. (2014) design a greedy-based algorithm called Greedy VM placement with Two Path Routing (GVMTPR) aimed at reducing the network traffic, the number of servers and resource utilization, with guaranteeing the protection. The algorithm splits flow into and route them through two link-disjoint paths to reduce congestion while satisfying the bandwidth and protection grade requirements. They use the maximum load on any link as a measure of congestion. Further, this scheme considers the network state for multipath routing. Traffic splitting helps to reduce congestion and ensure protection in the event of failures. Traffic flows with protection requirements can be split and sent across two or more paths. It is desirable to split large flows when compared to short flows. This is because large flows are more likely to cause link congestion. Further, they are small in number, implying lesser

Table 5
Objectives of Game algorithm, Graph theory and Greedy-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Xiao et al. (2015)		✓			
Kanagavelu et al. (2014)	✓				
Dong et al. (2013)		✓			
Breitgand et al. (2013)	✓				
Dai et al. (2014)		✓			

Table 4
Properties of bin packing and best fit decreasing-based VM placement schemes.

Schemes	SLA Support	Security	Environment	Algorithm type
Xiaoli and Zhanghui (2012)				Bin packing
Babu and Samuel (2014)	✓			Bin packing
Noumankhan Sayeedkhan and Balaji (2014)			IAAS	Best fit decreasing
Wang et al. (2014)	✓			Best fit decreasing
Calcavecchia et al. (2012)	✓		IAAS	Best fit decreasing

implementation overhead due to traffic splitting. The use of link-disjoint paths guarantees that at least one of two paths is available in single link failures. In this scheme, the VM placement and route selection module's function is to select the right set of servers for the VMs and the routes for communicating VMs. The chosen servers and routes must satisfy the server resource and network resource requirements while optimizing the use of resources. The details of VM to server mapping and paths with rates will be communicated to the module "VM and route deployment".

Some of the existing VM placement schemes try to optimize physical machines resources utilization or network resources utilization. However, in Dong et al. (2013) the authors propose a placement scheme to meet both of these factors. This scheme considers the network link capacity and physical server size including its CPU, memory, storage and bandwidth to reduce the network elements and number of active physical servers by sending them to the sleep state. In this scheme, the optimization of physical server by the VM placement is modeled as a bin packing problem (BPP), while the optimization of the network resources is modeled as quadratic assignment problem (QAP) both of which are NP-hard. The authors design a new method by combining hierarchical clustering with best fit (BF) to solve the multi-objective optimization, which, first, hierarchical clustering algorithm enables the VMs with large traffic to be placed on the same PM or the same access switch to reduce the network traffic and according to the clustering results, the BF algorithm is applied to optimize the PM resources.

The authors of Breitgand et al. (2013) investigate the issue of placing images and VM instances on the servers to increase the affinity between them to mitigate communication overhead and latency. By providing coordinated placements for VM instances and their images the number of local VM instances can be increased and the number of remote VM instances can be reduced.

They model this problem by extending the class constrained multiple knapsack problems (CCMK) and present a polynomial time local search algorithm for the same size images. This model focuses on an off-line placement problem, where the set of demands and available servers is known. To solve this issue, the local search algorithm can be applied as a basis for ongoing optimization which periodically improves the VM placement and greedy placement of a new set of VM instances by allowing migrations of the VMs.

In paper Dai et al. (2014) the authors propose to optimize the energy consumption in the data centers by intelligently placing the VMs in public clouds. They formulate the problem of minimizing the power consumption in the data centers as an integer programming problem and prove it to be NP-hard. They present two greedy approximation algorithms named the MinPow which attempts to decrease the power consumption by avoiding powering up unnecessary servers and the MinCom which aims at packing one user's VMs as much as possible on the same server or under the same rack to decrease the networking energy cost.

Table 6 shows the properties of Game algorithm-based, Graph theory-based and Greedy-based VM placement schemes. As indicated in this table, only few VM placement schemes consider the cloud security in their placement operations and decisions. Also, these schemes are proposed for various cloud environments.

Table 6
Properties of Game algorithm, Graph theory and Greedy-based VM placement schemes.

Schemes	SLA support	Security	Environment	Algorithm type
Xiao et al. (2015)	✓		Homogeneous	Game algorithm
Abazari and Analoui (2014)		✓		Graph theory
Kanagavelu et al. (2014)				Greedy
Dong et al. (2013)	✓		IAAS	Greedy
Breitgand et al. (2013)			Private cloud	Greedy
Dai et al. (2014)			Public cloud	Greedy

3.9. Banker algorithm-based VM placement

To prevent deadlock, while placing the VMs in the data center, some providers or brokers may consider that the servers are overloaded and the system is out of resources. For example, in Singh and Hemalatha (2013) authors present a VM placement solution which focuses on the deadlock avoidance using the banker algorithm and checks whether the system is safe state or unsafe state while the VM allocation. Moreover, this scheme tries to control and minimize the energy consumption and VM migrations in the cloud data center by not compromising the SLA. The proposed work runs with a different overload and a VM selection algorithm to check which suits the best.

In Piao and Yan (2010) authors propose a network aware VM placement and migration scheme which places the VMs on servers by considering the network conditions between the servers and the data storage. In addition, it reduces the data transfer time and considers scenarios in which instable network conditions mitigate the application performance, and deals with such scenarios by migrating the VMs to other servers. However, it does not provide the enforcement of the SLA because the optimized data access might still go over the time requirement in the SLA. In some extreme cases, the data access time might be intolerable for the application. Eventually, this decline of the network status may jeopardize the SLA between the cloud service provider and service user. In this scheme, the SLA only specifies the execution time for an application to complete processing its related data and migration is triggered when the execution time crosses the threshold that is specified in the SLA.

3.10. Heuristic-based VM placement

To address the problem of the online VM placement in Li et al. (2013) the authors propose an online algorithm called EAGLE, which is based on a multi-dimensional space partition model and minimizes the number of the running PMs and the total energy consumption. The idea of EAGLE is to make a trade-off between balancing multi-dimensional resource utilization and minimizing the local number of the PMs when placing the VMs at each time-slot. In this scheme, the space is divided into three domains which each dimension of the space, corresponds to a one-dimensional resource. When a new VM placement task arrives, EAGLE checks the posterior resource usage state for each feasible PM, and then chooses the most suitable PM. The basic operation is starting up a new PM to avoid excessive resource fragments. It is efficient in decreasing resource fragments over the long run, and reducing the number of the PMs.

In Fang et al. (2013) the authors present a novel approach, called VM Planner, for power reduction in the data centers, which applies a network-wide power manager to optimize both VM placement and traffic flow routing by turning off the network elements. They solve the placement problem by mapping VM-groups to server-racks on a one-to-one basis instead of mapping the VMs to individual servers on an N-to-1 basis. The VM-groups are obtained through the BMKP algorithm which ensures that VM pairs with a high mutual traffic rate are within the same VM-group. Then, mapping of VM-groups to server racks are realized by adopting the QAP algorithm which ensures that the traffic

Table 7
Objectives of Heuristic-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Vu and Hwang (2014)	✓	✓			
Li et al. (2013)		✓			
Fang et al. (2013)		✓			
Lucas Simarro et al. (2011)			✓		
Huang et al. (2013)		✓			
Su et al.	✓	✓			
Chen et al. (2012)		✓		✓	
Gupta et al. (2013)					✓
Goudarzi and Pedram (2012)		✓			
Duong-Ba et al. (2014)		✓			
Zhuang and Guo (2013)			✓		
Machida et al. (2010)				✓	
Fang et al. (2013)		✓			
Elmroth and Larsson (2009)	✓				
Somani et al. (2012)				✓	
Yapicioglu and Oktug (2013)	✓				
Liang et al. (2014)		✓		✓	
Le et al. (2011)		✓	✓		

distribution among VM groups could be better aligned with the communication distance between them. Thus, VM-groups with a large mutual bandwidth usage are assigned to host servers in close proximity. By such VM placement, the aggregate rates of inter-rack traffic and the total traffic demands can be minimized. Finally, the MCFP algorithm (Heller et al., 2010) is used to optimize routing and find the minimum-power network subset that satisfies the traffic conditions.

In Lucas Simarro et al. (2011) the authors present a cloud broker for virtual infrastructures deployments for dynamic pricing multi-clouds. It is composed of cloud manager and a cloud scheduler which optimizes virtual cluster placements on the available clouds. This scheduler uses some variables such as average prices or cloud prices as trends for an optimal deployment and is based on a prediction model that considers the historical prices of cloud providers and predicts the best next hour deployment. Moreover, this scheduler decreases user's investment while maintaining a particular user-managed virtual cluster performance. To achieve this goal, they take advantage of price variability offered by cloud providers by moving part of the cluster to the cheapest placement. The cloud manager addresses the management and monitoring actions using the OpenNebula (ONE) virtual infrastructure manager which provides a uniform user interface to manage the VMs in a pool of distributed resources.

In Huang et al. (2013) the authors provide a multi-dimensional space partition model to characterize the resource usage states of the PMs. In this model, each dimension of the space corresponds to one dimensional resource, and the whole space is partitioned into three domains with distinctive features, which indicate the suitability of resource utilization for each placement task.

Based on multi-dimensional space partition model, they propose a VM placement algorithm, which can balance the utilization of multi-dimensional resources, reduce the number of the running PMs and mitigate the energy consumption. When a new VM placement task arrives, this algorithm checks the posterior resource usage state for each feasible PM, and then chooses the most suitable PM. This algorithm can make the resource be utilized in a more balanced manner and introduces less and smaller resource fragments and further, consumes less energy.

Most existing works consider the inter-VM relations without considering the heterogeneity of the cloud data centers. In practice, such data centers are often partitioned into logical groups for load balancing and specific services, cloud users always assign their VMs with specific PM requirements, which make the inter-VM relations far more complex. To mitigate this problem, in Su et al. authors consider affinity and conflict between the VMs in the VM placement problem. They analyze the affinity and conflict between the VMs in the data centers and demonstrate the benefits of accounting for the affinity and conflict in the placement of the VMs.

In Gupta et al. (2013) authors focus on the problem of placement of multiple VM instances comprising a single job request to PMs by meeting the SLA requirements and improving resource utilization. For this purpose, a high performance computing-aware scheduling algorithm is designed for the VM placement which achieves better resource utilization and limits cross-application interference through careful co-location. This application-aware scheme allocates n VM instances to physical hosts from a single pool, and it is implemented on top of Open Stack Compute (Nova), and also incorporates it in a simulator (CloudSim). Table 7 shows the objectives of Heuristic-based VM placement schemes.

Goudarzi and Pedram (2012) present an approach to generate multiple copies of the VMs without sacrificing the QoS. They provide a heuristic algorithm based on dynamic programming and local search to determine the number of VM copies, and then place them on the servers. The proposed solution provides a flexible method to increase the energy efficiency of the cloud computing system or increases the resource availability in the data center. To guarantee QoS for each VM, only fixed memory bandwidth requirement is considered and a limitation on the number of VM copies is added.

In multiple geographically distributed data centers, it is better to place the VMs close to their users to achieve smaller response time, higher bandwidth and better performance. In Zhuang and Guo (2013) they propose a VM placement and assignment algorithm of OCPA in a large scale cloud deployment which runs quickly and can significantly reduce the network distance between the VMs and clients, hence improving the perceived user performance.

In Fang et al. (2013) they propose a power-efficient solution for the VM placement and migration in a fat tree data center network. This solution reduces power consumption as well as job delay by aggregating the VMs to a few hypervisors and migrating communicating parties to close locations. This scheme is based on fat tree topology and Open Flow protocol which is a software defined networking technology for customizing network protocols. In an Open Flow environment, a controller monitors job loads, VM requirements and resource availability, then Open Flow controller can schedule jobs and distribute the VMs accordingly. As jobs change and flows shift, the Open Flow controller dynamically adjusts VM assignments by placing the VMs near to each order to reduce energy consumption.

In VM placement schemes, the VMs are often consolidated to reduce the servers required to host them for reducing energy consumption, but this migration can increase the latency in some applications and may affect their performance. In Vu and Hwang (2014) authors present a VM placement solution called traffic and power-aware VM placement (TPVMP) which reduces power consumption and improves communication performance by reducing traffic cost of the VMs. Furthermore, in heterogeneous cloud data centers, this algorithm minimizes network congestion while energy consumption is unchanged. Also, it determines over-utilized PMs and selects the VMs to be migrated according to different migration policies. However, when the PM is under-utilized, it selects all hosted VMs and then finds a destination PM which mitigates communication cost among the VMs. They assume that traffic map inside a data center is stable and locate heavy communicating VMs on the near PMs. Also, on each PM, network-aware VMs are mixed with CPU-aware and CPU-network-balance ones to maximize resource utilization.

The authors of Somani et al. (2012) present a VM placement scheme which considers the resource contention between the VMs and their continuous usage. This scheme classifies the resource into three categories which are: CPU, network, and disk I/O. Then on the basis of resource usage patterns it differentiates between various VMs using a three dimensional vector called Resource Usage Vector (RUV) and applies a new VM placement algorithm. For this purpose, the VMs are rearranged according to their RUVs and those VMs which consume different resources are packed together to minimize resource contention. However, it will place two VMs using the same resources if one of the VMs uses the resource less extensively or normally. Also, they present a modified version of the same algorithm which takes care of the VMs with multiple VCPUs and restricts CPU over commitment. Both of these algorithms generate migration schedules which minimize the number of migrations by allocating a VM to its original host if all the constraints are satisfied.

In Yapicioglu and Oktug (2013) authors develop a traffic-aware VM placement algorithm which clusters frequently communicating VMs based on dynamic network traffic and places clusters into the racks of physical servers. This scheme minimizes networking delay based on the average communication path length, number of active servers and number of active network elements such as links and switches. Moreover, this scheme models the VMs as graph vertices which edge weight represents the cumulative network traffic flow between the VMs.

As network flow requests arrive, a network flow matrix is formed based on the flows in that period. Elements of this matrix contain network flow density information for the corresponding VM pairs. At the same time, target number of clusters is calculated which corresponds to the number of racks to be used. Initially, the VMs do not belong to a cluster and maximum number of clusters is defined equal to the number of the VMs. Clusters consist of computing slots, and cluster capacities are defined by the capacities of racks. This scheme follows a bottom-up approach to create or merge clusters of the VMs at each step. First, for each VM, traffic flows are sorted in a decreasing order. Then a pair of the VMs corresponding to the maximum flow is tried to be clustered together. Unless all the VMs are clustered and target number of clusters is reached at any step, all of the flows are processed. If the number of clusters is still higher than the target number of clusters after processing, they are sorted in a no decreasing number of occupied slots and the VMs in the clusters with the least number of occupied slots are merged into the other clusters until target number of clusters is reached. After clustering, inter-rack traffic is minimized. Next phase is to minimize overall network traffic at the core and intermediate links of the network. In this phase, the 'clusters of VMs' are clustered so that communication path between the VMs belonging to different clusters would be minimized. Once the racks of all the VMs are determined, the VMs are placed into the free computing slots sequentially.

Liang et al. (2014) present a reconfiguration framework based on request prediction to cope with the dynamic re-deployment of the VMs and resources in cloud data centers. Furthermore, to avoid the delayed reconfiguration, a prediction method based on Modifying Index Curve Model is presented. Also, an algorithm, called App-VM-Reconfiguration (AVMR), is developed to reconfigure the VMs and other resources. Based on the prediction of application request volumes, AVMR algorithm separates the configuration computation from the actual reconfiguration. It derives a specific reconfiguration in advance and avoids the potential delay between the varied requirement and the reconfiguration.

Table 8 exhibits some properties of Heuristic-based VM placement schemes. Again, the algorithm applied for VM placement, the simulation environment, the security and SLA support by each VM placement scheme is specified.

3.11. Automata-based VM placement

Rasouli et al. (2013) present a new approach based on the continuous action-set Learning Automata (CALA) for dynamic replacement of the VMs over heterogeneous data centers to reduce power consumption. This scheme supports live migration and forces idle nodes to sleep. Also, it tries to reduce energy consumption while maintaining service quality and mitigating the heat and greenhouse gases. For this purpose, those hosts which their CPU utilization is equal to zero are shut-downed. In this scheme, each server in the data center applies automata whose learning is accepting or rejecting the VM from the source node to the destination node. Also, each learning automata has two actions only one of which is active at a time. According to the utilization of servers, idle status, average status, active status and over utilized status are considered. Transition from one state to another is

Table 8
Properties of Heuristic-based VM placement schemes.

Schemes	SLA support	Security	Environment	Algorithm type
Vu and Hwang (2014)	✓		Heterogeneous	Heuristic
Li et al. (2013)				Heuristic
Fang et al. (2013)				Heuristic
Lucas Simarro et al. (2011)			Multicloud	Heuristic
Huang et al. (2013)				Heuristic
Su et al.		✓	Heterogeneous	Heuristic
Chen et al. (2012)				Heuristic
Gupta et al. (2013)	✓		IAAS	Heuristic
Goudarzi and Pedram (2012)	✓			Heuristic
Duong-Ba et al. (2014)				Heuristic
Zhuang and Guo (2013)				Heuristic
Machida et al. (2010)	✓			Heuristic
Fang et al. (2013)				Heuristic
Elmroth and Larsson (2009)	✓	✓	Federated cloud	Heuristic
Somani et al. (2012)			IAAS	Heuristic
Yapicioglu and Oktug (2013)				Heuristic
Liang et al. (2014)				Heuristic
Le et al. (2011)	✓			Heuristic

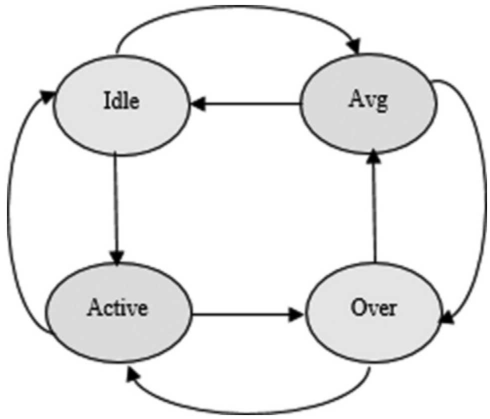

Fig. 9. Automaton state changes.

equivalent to the changing state of a node to another after the VM admission.

Figure 9 shows the general form of state change. Each learning automata has its own Markov chain and their coefficients of reward and penalty values are determined according to the source and destination nodes. Destination nodes use this environment response in the next iteration until the new assignment is created to be closer to the optimal schedule. These iterations recur until all automata are converged and identical allocation be created in recurrence. After that, finally VMs migrates to the host that accepts them.

3.12. Genetic algorithm-based VM placement

In Liu et al. (2014) authors propose a multi-objective VM placement model which minimizes the number of active PMs, also minimizes communication traffic and balances multi-dimensional resource usage within a data center. In this scheme, an improving evolutionary multi-objective algorithm called NS-GGA is designed which incorporates the fast no dominated sorting of NSGA-II into the Grouping Genetic Algorithms. The NS-GGA adopts fast non-dominated sort and grouping of genetic operators to approach the pareto-optimal front.

A VM placement model for maximizing resource utilization, balancing multi-dimensional resources usage and minimizing the data center traffic is proposed by Wang et al. (2013). In this scheme, the multi-objective placement problem is made simpler by applying average valued inequality and positional constraints. Also, an improved genetic algorithm is developed with local heuristic method and elitism strategy to adapt genetic operators with local heuristic method and Elitism strategy into iteration in finding the best placement.

In Yang et al. (2014) authors present a scheme called VM Placement and Traffic Configuration Algorithm or VPTCA. It is an energy-efficient data center network planning scheme that collectively deals with VM placement and communication traffic to reduce the data center network's energy consumption. To reduce traffic, in this scheme, communicating VMs are assigned to the physical machine. Moreover, VPTCA uses switch ports and link bandwidth for load balancing and avoiding congestions. This helps the data center network to increase its transmission capacity and save network energy. VPTCA takes these factors into account when making VM placement decisions in the GA. The succession of GA keeps the relationship and inherits to the offspring. In particular, they design a two-tier DNA coding method and a novel crossover operator to ensure that this inherent relationship is not to be harshly broken during the VM placement. Also, a dynamic routing table generation algorithm is proposed to save switches' temporary memory utilization even in the large-scale data center network traffic configuration.

The grouping genetic algorithm which generally is used for the VM placement, but it is not efficient always. For this purpose, in Jamali and Malektaji (2014) authors propose to improve grouping genetic algorithm by introducing an efficient technique for encoding and generating new solutions. By using the proposed crossover method, the complexity of managing the diverse set of the VMs as well as power consumption and resource wastage are reduced and significant cost reduction and operational efficiency are obtained. In this scheme, the problem of the VM placement is modeled using vector packing problem and it is tried to mitigate power usage and reduce the used servers and maximize resource usage.

In Xu and Fortes (2010) authors formulate the VM placement as a multi-objective optimization problem to simultaneously optimize conflicting objectives such as efficient usage of multi-dimensional resources, avoiding hotspots, and reducing energy consumption. For this purpose, fuzzy multi-objective evaluation is applied in the placement algorithm in order to combine the conflicting goals.

In this scheme, a two-level control system is proposed to manage the mappings of workloads to the VMs and the VMs to physical resources. Also, a modified genetic algorithm is proposed to effectively deal with the potential large solution space for the large-scale data centers. Table 9 specifies the objectives of genetic-based VM placement schemes. As shown in this table, some of these scheme pursue multiple objective to be more effective.

3.13. PSO-based VM placement

Some of the proposed VM placement solution utilize Particle Swarm Optimization (PSO) algorithm because of its faster converge speed. The main advantage of PSO is that each particle has an inbuilt memory and can keep track of its coordinates in the problem space associated with the optimal solution it has achieved so far. Moreover, it combines the local search methods with the global search methods to balance the exploration and exploitation.

In Dashti and Rahmani (2015) authors propose a PSO-based approach to reallocate the migrated VMs in the overloaded hosts and dynamically consolidating the under-loaded hosts for power saving. This scheme applies the PSO algorithm and maps one VM to each dimension of the particle's position where the value of the position indicates the power of host plus the cost of migration. The objectives of this scheme are energy efficiency to save money and maintaining QoS in the private cloud.

Wang et al. (2013) focus on energy consumption of the data centers and present a model for it. They present a VM placement scheme which tries to optimize energy consumption based on the PSO algorithm. They improve PSO by redefining its parameters,

Table 9
Objectives of Genetic-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Mi et al. (2010)		✓		✓	
Liu et al. (2014)	✓	✓			
Wang et al. (2013)	✓			✓	
Yang et al. (2014)		✓			
Jamali and Malektaji (2014)		✓			
Xu and Fortes (2010)				✓	

adopt an energy-aware local fitness first strategy to update the particle position and apply a two-dimensional particle encoding scheme. Finally, the improved PSO is used to provide optimal VM replacement by considering the energy consumption.

3.14. ACO-based VM placement

To solve the VM placement problem, Gao et al. (2013) present an ant colony system algorithm which achieves a set of non-dominated solutions to improve resource utilization and power consumption. The problem of the VM placement across a pool of server nodes is modeled as the multidimensional vector packing problems which dimensions are resource utilizations. This scheme uses two dimensions to characterize a VM and a server node which are CPU and memory usage. If two VMs are running on the same server, the CPU utilization of the server is estimated as the sum of the CPU utilizations of the two VMs. To prevent full CPU and memory usage, an upper bound on resource utilization of a single server is imposed with some threshold value. The main idea behind this is that full utilization can cause severe performance degradation and VM live migration technology consumes some amount of CPU processing capability on the migrating node. The results of this scheme is compared with multi-objective genetic algorithm, two single-objective algorithms, a bin packing algorithm and a max–min ant system algorithm.

In Dong et al. (2014) authors present an Ant Colony Optimization-based (ACO-based) VM placement solution and consider multi-resource constraints of the PM to improve network performance and optimize the total network traffic for more scalability. By minimizing the total traffic using the VM Placement in the data center and putting the VMs with large traffic on the same PM or on the same switch, they can modify the traffic layout between the VMs. By minimizing network Maximum Link Utilization (MLU), the network traffic is allocated evenly and congestion hotspots are avoided. Furthermore, to improve performance, they combine the ACO with 2-opt local search algorithm, because for large data calculation, ACO has a long way-finding, slow convergence and high time complexity.

The authors Ma et al. (2012) present a solution for the VM placement in IaaS cloud and consider the initial VM placement as a multi-objective optimization problem based on the ACO. This scheme uses the positive feedback mechanism and by pheromone updating, it achieves the optimal solution by the efficient convergence.

Moreover, as indicated in Fig. 10, this scheme is comprised of local and global managers (VMM) which the Local managers continuously monitor server's CPU utilization and try to manage the VMs by their resource requirements. Moreover, the global manager is positioned on a master node and to control the VM placement, it collects information from the local managers to achieve the overall view of the resources utilization. In this scheme, the VM size is considered as a D-dimensional vector which each dimension corresponds to one type of the requested resources. Resources on physical servers are allocated as slices along multiple dimensions according to the resource demands of the VM requests and each VM is assigned and bounded to a slice of a server. The monitoring system of the data center measures resource and power usage and QoS of each server to collect them into a centralized repository to be utilized by the global manager for the VM placement. Finally, global manager takes the VM placement decision by considering factors such as the SLA, resource and power. Table 10 shows the objectives of the metaheuristic-based VM placement schemes.

3.15. Simulated Annealing-based VM placement

Simulated Annealing or the SA is a probabilistic method to find a good and not necessarily perfect solution to an NP complete optimization problem. In Wu et al. (2012) authors propose the SAVMP or Simulated Annealing VM Placement algorithm to improve energy efficiency in the VM placement problem. The SAVMP applies the First Fit Decreasing method to generate an initial assignment of the VMs to the PMs. However, similar to the limitation of the FFD, it is not appropriate for the dynamic VM consolidation and it should be used for the static VM placement. In this scheme, the performance of the SAVMP is compared with the First Fit Decreasing and a multi-start random searching approach.

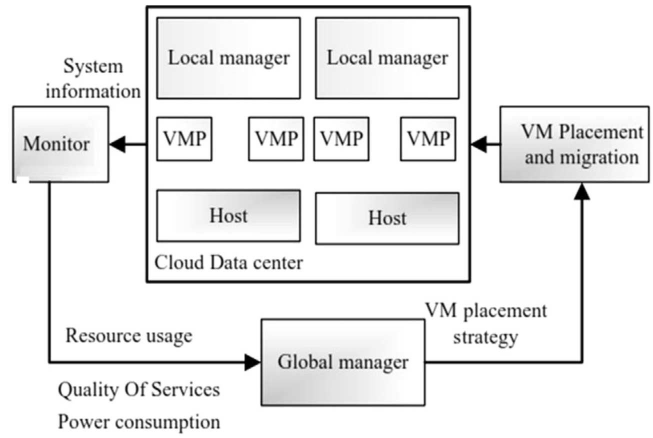

Fig. 10. System management structure.

3.16. Biogeography-based optimization

Biogeography-Based Optimization or BBO can guarantee convergence to the optimal solution. BBO applies biogeography to solve discrete optimization problems. Zheng et al. (2015) propose a novel multi-objective VM consolidated placement solution called VMPMBBO. This scheme treats the consolidated placement problem as a complex system and tries to minimize both the resource wastage and the power usage. They indicate that the VMPMBBO has better convergence and is more efficient. Moreover, they apply a resource wastage model which considers the CPU, RAM and bandwidth usage. Also, their power consumption model considers the CPU utilization of servers, and idle servers are turned off to prevent further power consumption.

3.17. Hybrid VM placement

Standard Genetic Algorithm (GA) evolves its solution probabilistically and this sometimes leads into a total random search which may be slow in finding a global optimal solution. On the other hand, with the PSO, it is faster in finding a solution but may face a problem of being trapped in a local optima solution. Similarly, Ant Colony Optimization (ACO) has a mixture of the problems faced in the Particle Swarm Optimization (PSO) and GA. The hybridized met heuristic algorithms take advantages of each combined algorithm. For example, in Mark et al. (2011) authors propose a hybrid solution called EOVMP algorithm by using the GA, PSO and ACO algorithms. This scheme applies a demand forecaster to obtain the near optimal VM placement and minimize the users' cost. As indicated in Fig. 11, the EOVMP uses predicted demand to allocate the VMs using reservation and on-demand plans for job processing. Moreover, prediction algorithms such as simple Kalman filter, double exponential smoothing, and Markov prediction are applied to forecast the resource from cloud users to optimize resource provisioning. In this scheme, cloud broker is responsible for provisioning the required amount of the VM resources to multiple cloud providers. The objective is to minimize the cost by predicting the demand of cloud users and optimizing

Table 10
Objectives of the PSO, Banker algorithm, ACO, BBO and Gray correlation degree-based VM placement schemes.

Schemes	Traffic-aware	Energy-aware	Cost-aware	Resource-aware	Application-aware
Dashti and Rahmani (2015)		✓	✓		
Singh and Hemalatha (2013)		✓			
Piao and Yan (2010)	✓				
Wang et al. (2013)		✓			
Gao et al. (2013)		✓		✓	
Dong et al. (2014)	✓				
Ma et al. (2012)		✓		✓	
Wu et al. (2012)		✓			
Zheng et al. (2015)		✓		✓	
Mark et al. (2011)				✓	
He (2014)		✓			

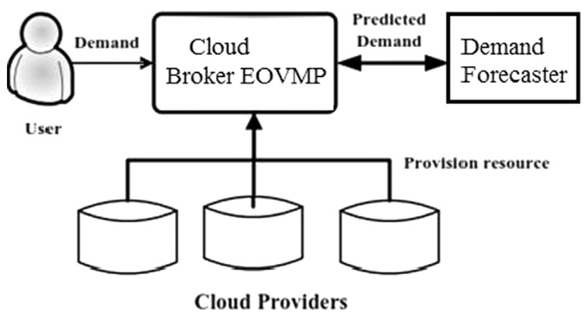

Fig. 11. Simple cloud computing environment.

the VM placement based on the user's usage history. The EOVMP improves the convergence speed by abstracting the message passing mechanism of the PSO or path marking similar to that in the ACO. In each iteration, the local best solution (pBest) and the global best solution (gBest) are tracked and stored.

These best solutions will then be propagated to the next iteration where the crossover/update of the solution is performed with the best solution. As a result, the solution can quickly converge to the global optimal solution. To prevent the solutions from being trapped in local optimal, mutation and the resetting of the population are performed. Mutation helps in the diversification of the solution by resetting the population. Therefore, the algorithm can explore the solution space more extensively.

Table 11 indicates the properties of the met heuristic-based VM placement schemes.

3.18. Correlation-based VM placement

The gray correlation theory can solve the problem of multi-objective decision-making in uncertain environments. The gray correlation degree is used to represent the similarity of an ideal solution and a candidate one. He (2014) presents a new multi-objective decision-making method of the VM placement based on gray correlation degree. This scheme analyzes the influences of the CPU utilization change on the SLA violation and energy consumption and the influences of the server load on the migration number of the VMs. Then, it focuses on the SLA violations of those servers of the VMs in migrating. For a physical server, the SLA violation is usually defined as the difference between the total MIPS requested by all the VMs from the total MIPS allocated to its VMs. When the CPU utilization of a physical server exceeds some threshold the SLA violation increases rapidly. This scheme builds three single-objective evaluation functions and establishes the multi-objective decision-making model of the VM placement. Furthermore, after the VM placement changes of the SLA violation, energy consumption, and the server load before the placement are selected as decision factors.

Resource sharing in virtualized data centers greatly helps to improve resource utilization, but such resource sharing has performance implications. To mitigate this problem, in Do et al. (2011) authors present a novel application profiling technique using the Canonical Correlation Analysis (CCA) method, which identifies the relationship between the application performance and resource usage. They also, present a performance prediction model based on application profiles generated using the CCA. The CCA has been widely used to find the linear relationship between the two sets of variables and is capable of modeling correlation pattern for predictions. In this scheme, a given application is executed with different background loads and the links between the application's performance and its resources utilization is achieved through the correlation analysis. The output of the correlation analysis is then used as the profile of the application which represents the level involvement of the system factors into the correlation with the application performance. The effective profiling of the applications is of great importance in various respects because the quality of consolidation decisions can be improved based on resource utilization and profit. Also, the application performance can be guaranteed.

4. Cloud simulators and providers

This section analyzes some of the simulators and cloud providers which have been utilized in evaluation and analysis of the VM placement schemes. Each subsection illustrates the capabilities and properties of each environment.

4.1. Amazon EC2

Amazon Elastic Compute Cloud or Amazon EC2 is a web service which provides scalable computing capacity in the cloud and is designed to make large-scale cloud computing easier for cloud developers. EC2 can be used to achieve the required number of virtual servers, configure security and networking, and manage storage. Also, it enables to scale up or down to deal with changes in requirement, decreasing the need to forecast traffic. Amazon EC2 mitigate the boot time of new server instances to few minutes and provides quick scale up and down capacity, the computing requirements are changed. Also, it provides developer tools to build the desired applications.

4.2. XGE

Xoreax Grid Engineer (XGE) technology converts a collection of computers into a private cloud where all systems help to provide a dynamic HPC environment, to lower maintenance costs. This platform allows applications to apply unutilized CPU on the network and can use the public clouds. IncrediBuild-XGE's architecture accelerates

Table 11
Properties of the PSO, Banker algorithm, ACO, BBO and correlation-based placement schemes.

Schemes	SLA support	Security	Environment	Algorithm type
Dashti and Rahmani (2015)			PAAS	PSO
Singh and Hemalatha (2013)	✓		IAAS	Banker algorithm
Piao and Yan (2010)	✓			Banker algorithm
Wang et al. (2013)				PSO
Gao et al. (2013)				Ant colony
Dong et al. (2014)				Ant colony
Ma et al. (2012)	✓		IAAS	Ant colony
Wu et al. (2012)				SAVMP
Zheng et al. (2015)				BBO
Mark et al. (2011)				ACO+PSO+Genetic
He (2014)	✓			Gray correlation degree

computational processes running on the MS Windows, allowing them to become "grid-enabled". Moreover, using IncrediBuild-XGE:

- Requires no resource management
- Eliminates the need for the dedicated hardware
- Ensures quick implementation
- Requires no virtualization image bank

4.3. XEN

The XEN Cloud Platform is an open source virtualization solution which includes the XEN Hypervisor, the enterprise ready XEN API tool stack and integrations for cloud, storage and networking solutions. XEN Cloud Platform provides services such as: support for windows and Linux guests, performance monitoring and alerting, upgrading, patching capabilities, live snapshots, live migration of VMs, flexible storage, networking and etc.

4.4. ClouDIA

ClouDIA is a cloud infrastructure which utilizes various virtualization technologies like XEN and considers the SLAs in its provided services. Figure 12 depicts the main components of ClouDIA and their communications with the public cloud (Zou et al., 2012).

Generally, the ClouDIA deploys the VMs in the public cloud in four steps. First, the public cloud tenant requests ClouDIA for a number of the VMs and the required VMs communication graph. ClouDIA over allocates instances to increase the chances of finding a good deployment. Then, the best deployment plan is searched using integer programming and the constraint programming. Finally, any over-allocated instances are terminated and the public cloud tenants may start the application with the deployment plan.

4.5. Comparison

This subsection presents a comparison between the simulator and environments having been applied in each scheme for simulation purpose. Figure 13 indicates the number of simulators, programming languages and cloud providers, applied in the proposed VM placement schemes for the cloud computing environment.

Also, Table 12 provides useful information about scenarios which have been considered in the simulations of VM placement schemes. The following items are highlighted in this table:

- Datacenter specifications
- PM specifications
- Network topology
- VMs specifications

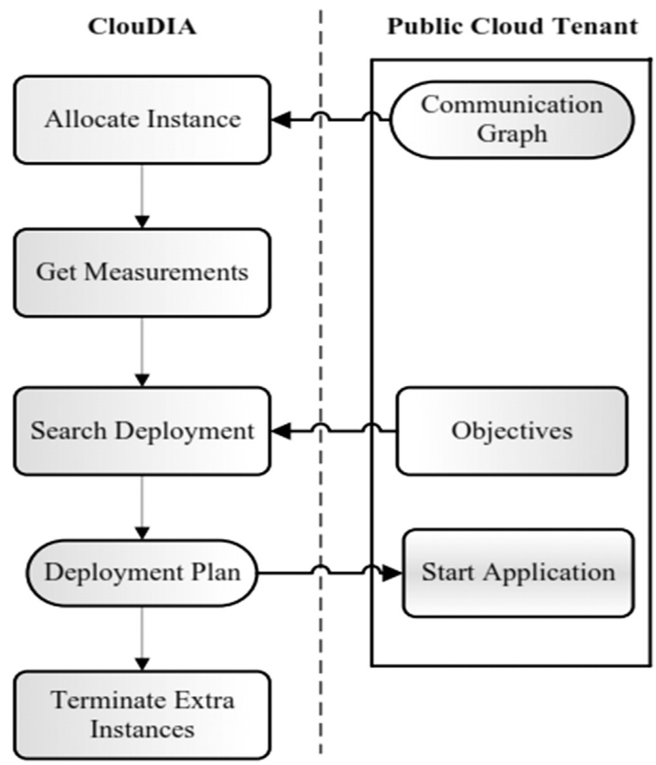

Fig. 12. ClouDIA architecture.

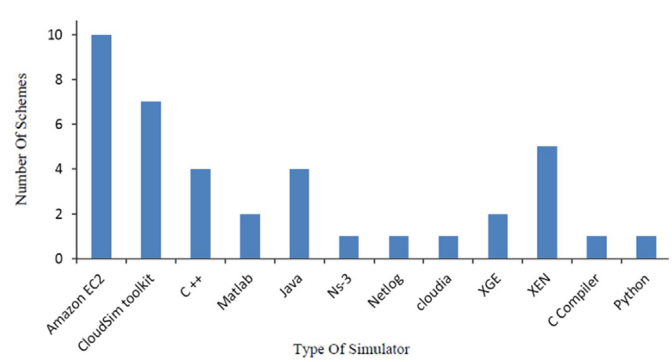

Fig. 13. Cloud simulators and providers applied in the VM placement schemes.

Each column of Table 12 represents a different aspect of applied scenarios. For example data center column shows the type or number of data center used in the simulations. Also, the PM specification column includes hosts in detail. Network topology column indicates the network topology which authors have considered in their scenarios and VMs specification column specifies

Table 12
Simulation scenarios in VM placement schemes.

Scheme	Datacenter specifications	PM specifications	Network topology	VMs specifications
Dong et al. (2014)	–	64 PMs	Fat-tree, Tree, VL2	256 VMs
Liang et al. (2014)	–	• 92 blade servers (with 736 Processor 2.6 GHz) • 1.5 TB memory	_	_
Le et al. (2011)	3 Data Center	• 1440 server (4 cores) • 4 GB memory	_	_
Kanagavelu et al. (2014)	_	128PM	Fat tree	• Large (18–22%) of total resource • Small VMs (1–1.5%) of resources
Jayasinghe et al. (2011)	_	• Small data centers, 60 PM • Large data centers, 200 PM	_	20–100 VMs
Dupont et al. (2012)	Two HP Data Center	CPU intel xenon E5520 2.27 GHz, 24 G	_	Each VM uses 70% PM CPU and 100 MB RAM
Zamanifar et al.	Two computing centers, Two storage centers	100 host	_	50 VM
Xiaoli and Zhanghui (2012)	_	_	_	64 VMs, 100 VMs, 128 VMs, 200 VMs
Babu and Samuel (2014)	_	10 PM, each has 1–3 GB memory	_	100 VMs
Noumankhan Sayeedkhan and Balaji (2014)	_	10 hosts, each contains one core CPU, 5 GB RAM	_	20 VMs, RAM 512 MB, Bandwidth 2500 MB
Wang et al. (2014)	_	_	• 3 tier fat-tree network, • 16 core-level, 32 aggregate level, 32 edge-level switches for 8 PM with 4VMs.	256 VMs
Xiao et al. (2015)	_	• Dell Precision T1700 (4 processors) • 8 GB memory, • 2 TB SATA hard disk • 1 Gbit Ethernet interface	_	• 10, 16–25 VMs 1. CPU (1-2-3-4), 2. RAM (1 gig and 2 gig)
Abazari and Analoui (2014)	_	81 Host	_	100 VMS
Breitgand et al. (2013)	_	137 PM, 250 GB storage	_	1984 VMs
Dai et al. (2014)	_	• 100 host, cores (1,2,4,8) • RAM (3.75,7.5,15,30) • hard disk (4,32,80,160)	_	10–30 VMs
Vu and Hwang (2014)	_	150 PMs	• 1 core switch • 3 aggregation switches each is connected 5 edge switches	300–400 VMs
Fang et al. (2013)	_	2000 host	_	200 VMs
Lucas Simarro et al. (2011)	_		_	10 VMs
Huang et al. (2013)	_	• (12 GB or 24 GB) memory • (12 or 24) CPU cores	_	60–100 VMs
Su et al.	_	_	_	20, 50, 100, 150, 200 VMs
Chen et al. (2012)	_	Physical server's TCV {16*2.4 GHZ, 24 GB, 100 M}	_	• CPU:[.25,4]*2.4 • Memory:{0.5, 1, 1.5, 2.0, 3.0, 4.0, 6.0} • network I/O:{4, 6, 10, 15, 20}
Goudarzi and Pedram (2012)	_	Intel processors	_	• 200–5000 VMs • CPU(1 and 18 billion cycles per second) • Memory BW between 768 MB/s and 4 GB/s.
Duong-Ba et al. (2014)	_	• Small(CPU=1 core, memory 3.75 GB HDD=4 GB), • Medium(CPU=2 core, memory 7.5 GB HDD= GB) • Large(CPU=4 core, memory 15 GB HDD=40 GB • Xlarge (CPU=8 core, memory 30 GB HDD=80 GB	_	250 random VMs
Fang et al. (2013)	_	• 128 PMs, 1024 PMs	_	20 VMS
Mi et al. (2010)	_	• 300 heterogeneous PMs • 100 Dells with Intel(R) Core(TM)2 Duo 2.83 GHz • 100 AMD Athlon(tm) 64 × 2 3600+1.9 GHz.	_	3 VMs
Jamali and Malektaji (2014)	_	[18,104] PMs	_	120 and 250 VMs
Xu and Fortes (2010)	_	128 PMs from 20 different type	_	250 VMs, CPU {0.25 0.5 1 1.5 2 2.5 3 4}GHz, memory {0.25 0.5 1 1.5 2 2.5 3 4} GB
Piao and Yan (2010)	_	3 host	_	3 VMs
Wang et al. (2013)	_	Intel core2, CPU 2.1 GHz, 2.0 G RAM	1000 cluster, each contains 350 PM	_

Gao et al. (2013)	–	Intel Pentium dual-core processor with 2.50 GHz, 3 G RAM	–	200 VMs
Ma et al. (2012)	–	100 PM, CPU core with 1000, 2000 or 3000 MIPS, 10 G RAM, 1 G bandwidth, 1 Tb storage	–	230 heterogeneous VMs
Mark et al. (2011)	–	• Intel Core 2, 2 GB RAM.	–	–
He (2014)	–	• 2000 or 3000 CPU capacity in MIPS • 4096 memory • 1 Gbit/s bandwidth	–	150 VMs: 1. CPU: 500,750,1000 or 1500 MIPS 2. Memory: 600,1750,850 3. Bandwidth 100 Mbps

the type or number of VMs that have been applied in the experiments.

5. Discussion

This section provides a complete comparative analysis of the VM placement schemes proposed for the cloud environment. It mainly provides a comparison of the following issues:

- The number of schemes provided for each algorithm.
- The VM placement factors applied in each scheme.
- The resources which have been considered in each.
- The power-related factors in the VM placement schemes.
- The cost factors in the VM placement schemes.
- The network factors in the VM placement schemes.

Figure 14 indicates the number of the VM placement schemes presented based on each algorithm.

5.1. Resource-aware VM placement schemes

Virtual resource management is one of the important issues in cloud computing. Normally, each VM may access various resources to execute its applications. Resource-aware VM placement schemes consider hardware resource requirements of the VMs in the placement decisions. An efficient resource-aware placement scheme tries to optimally place VMs on the PMs such that the overall resource utilization is maximized. Resource contention amongst multiple co-hosted neighboring VMs form the basis of the presented novel approach. Table 13 indicates the placement factors which resource-aware VM placement schemes consider in the VM placement process. This table's fields are Memory, I/O device, Ram, CPU and Network (Gao et al., 2013). To provide better placement results, resource-aware schemes often need some information about the various resources' state in data centers which can be achieved from the local agents in each physical machine. As it is shown in this table, CPU is the mostly considered factor in the resource-aware VM placement schemes.

5.2. Power-aware VM placement schemes

Due to the high cost of power consumption as well as the concerns for global warming and CO_2 emissions, recently, the green data centers have become an important issue for cloud service providers.

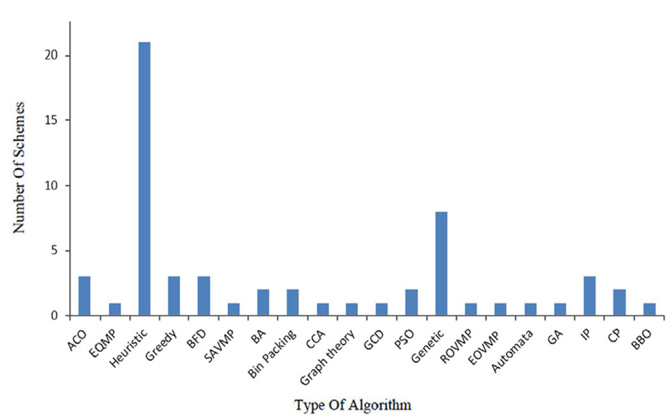

Fig. 14. Algorithms applied in the VMP schemes in cloud computing. BA: Banker Algorithm, CP: Constraint Programming, IP: Integer Programming, BFD: Best Fit Decreasing, BBO: Biogeography-Based Optimization, GA: Game Algorithm, CCA: Canonical Correlation Analysis, and GCD: Gray Correlation Degree.

Table 13
Comparison of resource-aware VM placement schemes.

Schemes	CPU	Memory	I/O device	Ram	Network	Disk I/O	Reduce active PM
Do et al. (2011)	✓	✓	✓				
Ma et al. (2012)	✓			✓			
Chen et al. (2012)	✓	✓			✓		
Noumankhan Sayeedkhan and Balaji (2014)	✓			✓	✓		
Gao and Tang (2013)	✓	✓					
Liang et al. (2014)	✓	✓			✓		
Gao et al. (2013)	✓	✓					
Calcavecchia et al. (2012)	✓						
Lin and Chen (2012)	✓	✓					
Somani et al. (2012)	✓				✓	✓	
Mi et al. (2010)							✓
Babu and Samuel (2014)							✓
Wang et al. (2013)	✓	✓			✓		
Xu and Fortes (2010)	✓	✓					
Machida et al. (2010)							✓
Zheng et al. (2015)	✓	✓			✓		

Table 14
Comparison of power-aware VM placement schemes.

Schemes	CPU utilization	Server power usage	PM states	Network elements	Data center power usage	Distance between VMs
Ma et al. (2012)	✓					
Xiao et al. (2015)			✓			
Rasouli et al. (2013)	✓					
He (2014)	✓					
Vu and Hwang (2014)			✓			
Dupont et al. (2012)	✓		✓			
Wang et al. (2014)		✓				
Li et al. (2013)	✓					
Fang et al. (2013)				✓		
Liang et al. (2014)		✓				
Le et al. (2011)					✓	
Zamanifar et al.		✓				
Singh and Hemalatha (2013)	✓					
Xiaoli and Zhanghui (2012)	✓					
Gao et al. (2013)	✓					
Babu and Samuel (2014)			✓			
Mi et al. (2010)			✓			
Liu et al. (2014)			✓			
Wu et al. (2012)	✓					
Su et al.			✓			
Huang et al. (2013)	✓		✓			
Chen et al. (2012)		✓				✓
Yang et al. (2014)				✓		
Goudarzi and Pedram (2012)			✓			
Dong et al. (2013)		✓		✓		
Jamali and Malektaji (2014)		✓	✓			
Duong-Ba et al. (2014)					✓	
Wang et al. (2013)		✓				
Fang et al. (2013)			✓			
Zheng et al. (2015)		✓		✓	✓	
Dai et al. (2014)		✓		✓		

Power-aware VM placement schemes are very important in green cloud computing which tries to make cloud data centers more efficient and reduce the power consumption of the data centers. This, finally results in lower maintenance costs for the cloud provider. Numerous power-aware VM placement schemes are proposed for cloud computing environment. For this purpose, they reduce the number of active, server, networking or other data centers' components.

- Avoiding powering-up unnecessary servers
- VM consolidation
- Packing VMs on the same server, under the same rack and so on.
- Switching VMs on and off
- Turning off the network's elements

Table 14 indicates the factors which power-aware VM placement schemes consider in their placement decisions. This table's fields are as follows:

- CPU Utilization: The state of being idle or busy according to the utilization of processors, we have defined some status. Four defined status explain as Idle, Average, Active, Over utilized (Rasouli et al., 2013)
- Server states: The amount of power that various server components (CPU, memory, etc.) consume
- PM states: reducing turned on pm, different states of pm consume different amounts of power, we have 4 states (running, Ready, Sleep, Off) (Xiao et al., 2015)

Table 15
Comparison of network-aware VM placement schemes.

Schemes	VMs cost	Physical machine cost	Cooling cost	Data center cost	Distance between VMs and clients
Chaisiri et al. (2009)	✓				
Le et al. (2011)			✓	✓	
Lucas Simarro et al. (2011)	✓				
Zhang et al. (2014)		✓			
Tordsson et al. (2012)		✓			
Mark et al. (2011)		✓			
Chen et al. (2013)	✓				
Zhuang and Guo (2013)					✓

Table 16
Comparison of cost-aware VM placement schemes.

Schemes	Minimizing the data transfer time	Traffic between VMs	Traffic between remote cloud sites	Traffic between PMs	Communication distance between VMs
Yapicioglu and Oktug (2013)		✓			
Vu and Hwang (2014)		✓			
Piao and Yan (2010)	✓				
Kanagavelu et al. (2014)		✓			
Dong et al. (2014)		✓			
Jayasinghe et al. (2011)		✓			
Liu et al. (2014)				✓	
Su et al.				✓	
Wang et al. (2013)				✓	
Meng et al. (2010)					✓
Breitgand et al. (2013)		✓			
Zhang et al. (2014)		✓			
Elmroth and Larsson (2009)		✓			

- Network elements: Reducing network routing elements' costs.
- Data center power usage: The cost is divided into four categories: server base energy, server dynamic energy, cooling energy and peak power.
- Distance between the VMs: decreasing distance between the VMs to save energy

5.3. *Network-aware VM placement schemes*

The intra-data center and inter-data center network traffic has a significant effect on the SLAs, revenue of the cloud providers and the performance of the cloud computing services. With the increasing trend towards communication applications, network-aware VM placement schemes have become very important. These schemes address the traffic related issues in the VM placement for the cloud data centers and try to reduce network traffic. Also, some of these schemes distribute network traffic evenly and try to avoid congestion. Generally, the VMs make use of communication network because of the following reasons:

- Data-intensive applications, which the VMs use, network to achieve its required data from the storage components.
- Communication with other VMs, for example, the VM may execute a task which is dependent on the other workflow tasks and it should achieve some data from its predecessor tasks.

To reduce the inter-data center and intra-datacenter traffic, network-aware VM placement schemes allocate the more communicating VMs to the same server, the same switch, the same rack or in the federated cloud on the same data center. Table 15 indicates the factors which network-aware VM placement schemes apply to produce placements that will result in less communication in the data centers. This table's fields are explained as follows:

- Minimizing the data transfer time consumption: Optimizing the data transfer between the virtual machine and data.
- Traffic between the VMs: Decreasing the overall network traffic and minimizing the average path length between the VMs, improve communication performance by reducing, minimizing the congestion of network between the VMs.

- Traffic between remote cloud sites: Reducing traffic in the federated clouds.
- Traffic between the PMs
- Distance between the VMs

5.4. *Cost-aware VM placement*

Cost-aware VM placement schemes try to mitigate the data centers maintenance costs for the cloud providers. The important issue in these schemes is to achieve the cost saving while considering the QoS of cloud services and honoring the SLAs.

Table 16 indicates the factors applied in the cost-aware VM placement schemes to produce a better placement. This table's fields are explained as follows:

- VMs cost: The cost of powering the VMs when they are instantiated in the data centers

- Physical machine cost: The cost of using the PM in the specific period of time
- Data center cost
- Distance between the VMs and the clients: By reducing the network distance between the VMs and the clients the perceived user performance is improved.
- Cooling cost: The cost of the cooling system that is used for the data centers

6. Conclusion

Cloud computing is an emerging computing platform which presents virtual resources pool by centralizing computing resources connected with network and presents them as various services such as IaaS, PaaS, SaaS. Virtualization is one of the necessary technologies which make the cloud computing possible. Recently, the VM placement problem has become a hot topic in the cloud data centers because proper placement can greatly reduce the VMs traffic in the data center and improve the physical machines' resource utilization. Also, effective VM placement algorithms can reduce the power consumption of the data centers by consolidating the VMs into fewer physical machines.

This paper presents a comprehensive survey and analysis of the VM placement schemes proposed for the cloud computing data centers. For this purpose, first the definition and objectives of the VM placement are provided, and then a classification of the VM placement schemes is presented based on the type of the placement algorithm. Then various proposed VM placement schemes are described and the VM placement factors in each scheme are investigated to illuminate the advantages, properties and limitations of each placement scheme in detail. Moreover, complete comparative comparisons of the VM placement schemes are presented which highlight the items that should be considered in the future researches and studies. As specified in these comparisons, most VM placement schemes are aimed to improve the performance and energy-related issues in the cloud systems and have neglected the security related objectives in the VM placement operations. But, with the ever increasing security attacks on the VMs and the cloud services, providing security in the cloud environment has become a more critical issue. For example, many attacks such as VM Escape attacks, VM Sprawling attacks, Cloud-Internal Denial of Service attacks (CIDoS), VM Neighbor attacks etc. are conducted in the cloud computing that target the VMs and a majority of them force the system to migrate the VMs. By misusing the VM migration feature of the cloud, these attacks incur high overheads to the cloud system and degrade its performance and availability. As a result, security is one of the crucial factors which should be considered in the future VM placement researches and studies, and its negligence makes the whole cloud system vulnerable to various security attacks and hinder the cloud computing critical role in the future IT systems.

References

Abazari F, Analoui M. Exploring the effects of virtual machine placement on the transmission of infections in cloud. In: Proceedings of the 7th International Symposium on Telecommunications (IST), 2014. IEEE.

Buyya R, et al. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener Comput Syst 2009;25(6):599–616.

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

Babu K, Samuel P. Virtual machine placement for improved quality in IaaS cloud. In: Fourth International Conference on Advances in Computing and Communications (ICACC), 2014. IEEE.

Breitgand D, et al. Network aware virtual machine and image placement in a cloud. In: 2013 9th International Conference on Network and Service Management (CNSM), 2013. IEEE.

Chaisiri S, Lee B-S, Niyato D. Optimal virtual machine placement across multiple cloud providers. In: IEEE Asia-Pacific Services Computing Conference, 2009 (APSCC 2009). IEEE.

Calcavecchia NM, et al. VM placement strategies for cloud scenarios. In: IEEE 5th international conference on Cloud Computing (CLOUD), 2012. IEEE.

Chen W, et al. A profit-aware virtual machine deployment optimization framework for cloud platform providers. In: IEEE 5th international conference on Cloud Computing (CLOUD), 2012. IEEE.

Chen K-Y, et al. Intelligent virtual machine placement for cost efficiency in geo-distributed cloud systems. In: IEEE International Conference on Communications (ICC), 2013. IEEE.

Davidovic V, et al. Private cloud computing and delegation of control. Procedia Eng 2015;100:196–205.

Dashti SE, Rahmani AM. Dynamic VMs placement for energy efficiency by PSO in cloud computing. J Exp Theor Artif Intell 2015:1–16.

Dupont C, et al. An energy aware framework for virtual machine placement in cloud federated data centres. In: Third international conference on future energy systems: where energy, computing and communication meet (e-Energy), 2012. IEEE.

Dong J, et al. Virtual machine placement for improving energy efficiency and network performance in IaaS cloud. In: IEEE 33rd International Conference on Distributed Computing Systems Workshops (ICDCSW), 2013. IEEE.

Dong J, et al. Energy-saving virtual machine placement in cloud data centers. In: 13th IEEE/ACM international symposium on Cluster, Cloud and Grid Computing (CCGrid), 2013. IEEE.

Dai X, Wang JM, Bensaou B. Energy-efficient virtual machine placement in data centers with heterogeneous requirements. In: IEEE 3rd International Conference on Cloud Networking (CloudNet), 2014. IEEE.

Duong-Ba T, et al. Joint virtual machine placement and migration scheme for datacenters. In: Global Communications Conference (GLOBECOM), 2014 IEEE. IEEE.

Dong J-K, et al. Virtual machine placement optimizing to improve network performance in cloud data centers. J China Univ Posts Telecommun 2014;21(3):62–70.

Do AV, et al. Profiling applications for virtual machine placement in clouds. In: IEEE international conference on Cloud Computing (CLOUD), 2011. IEEE.

Elmroth E, Larsson L. Interfaces for placement, migration, and monitoring of virtual machines in federated clouds. In: Proceedings of eighth international conference on Grid and Cooperative Computing, 2009. GCC'09. IEEE.

Fang W, et al. VMPlanner: optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers. Comput Netw 2013;57(1):179–96.

Fang S, et al. Power-efficient virtual machine placement and migration in data centers. In: IEEE International Conference on Green Computing and Communications (GreenCom), IEEE Cyber, Physical and Social Computing, 2013 IEEE and Internet of Things (iThings/CPSCom), 2013. IEEE.

Gahlawat M, Sharma P. Survey of virtual machine placement in federated Clouds. In: IEEE International Advance Computing Conference (IACC), 2014. IEEE.

Gupta RK, Pateriya R. Survey on virtual machine placement techniques in cloud computing environment. Int J Cloud Comput: Serv Archit 2014;4(4).

Gupta A, et al. Hpc-aware vm placement in infrastructure clouds. In: IEEE International Conference on Cloud Engineering (IC2E), 2013. IEEE.

Goudarzi H, Pedram M. Energy-efficient virtual machine replication and placement in a cloud computing system. In: 5th International conference on Cloud Computing (CLOUD), 2012 IEEE.

Gao Y, et al. A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. J Comput Syst Sci 2013;79(8):1230–42.

Gao J, Tang G. Virtual machine placement strategy research. In: Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on. 2013. IEEE.

B. Heller et al. ElasticTree: saving energy in data center networks. In: Proceedings of NSDI; 2010.

Huang W, Li X, Qian Z. An energy efficient virtual machine placement algorithm with balanced resource utilization. In: Seventh international conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013. IEEE.

He L. A method of virtual machine placement based on gray correlation degree. In: 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2014. IEEE.

Jayasinghe D, et al. Improving performance and availability of services hosted on iaas clouds with structural constraint-aware virtual machine placement. In: IEEE International Conference on Services Computing (SCC), 2011. IEEE.

Jamali S, Malektaji S. Improving grouping genetic algorithm for virtual machine placement in cloud data centers. In: 4th International eConference on Computer and Knowledge Engineering (ICCKE), 2014. IEEE.

Komu M, et al. Secure Networking for Virtual Machines in the Cloud. In: IEEE International Conference on Cluster Computing Workshops (CLUSTER WORKSHOPS), 2012. IEEE.

Kanagavelu R, et al. Virtual machine placement with two-path traffic routing for reduced congestion in data center networks. Comput Commun 2014;53:1–12.

Lombardi F, Di Pietro R. Secure virtualization for cloud computing. J Netw Comput Appl 2011;34(4):1113–22.

Li W, Tordsson J, Elmroth E. Virtual machine placement for predictable and time-constrained peak loads. In: Economics of grids, clouds, systems, and services. Springer; 120–34.

Li X, et al. Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center. Math Comput Model 2013;**58** (5):1222–35.

Lucas Simarro JL, et al. Dynamic placement of virtual machines for cost optimization in multi-cloud environments. In: International conference on High Performance Computing and Simulation (HPCS), 2011. IEEE.

Liang Q, et al. The placement method of resources and applications based on request prediction in cloud data center. Inf Sci 2014;279:735–45.

Le K, et al. Reducing electricity cost through virtual machine placement in high performance computing clouds. In: Proceedings of 2011 international conference for high performance computing, networking, storage and analysis. 2011. ACM.

Liu C, et al. A new evolutionary multi-objective algorithm to virtual machine placement in virtualized data center. In: 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2014. IEEE.
Lin J-W, Chen C-H. Interference-aware virtual machine placement in cloud computing systems. In: International Conference on Computer & Information Science (ICCIS), 2012. IEEE.
Machida F, Kawato M, Maeno Y. Redundant virtual machine placement for fault-tolerant consolidated server clusters. In: Network Operations and Management Symposium (NOMS), 2010 IEEE. IEEE.
Mi H, et al. Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers. In: IEEE international conference on Services Computing (SCC), 2010. IEEE.
Ma F, Liu F, Liu Z. Multi-objective optimization for initial virtual machine placement in cloud data center. J Inf Comput Sci 2012;**9**(16).
Mark CCT, Niyato D, Chen-Khong T. Evolutionary optimal virtual machine placement and demand forecaster for cloud computing. In: International conference on Advanced Information Networking and Applications (AINA), 2011 IEEE. IEEE.
Meng X, Pappas V, Zhang L. Improving the scalability of data center networks with traffic-aware virtual machine placement. In: Proceedings IEEE INFOCOM, 2010. IEEE.
Noumankhan Sayeedkhan P, Balaji S. Virtual machine placement based on disk I/O load in cloud. Int J Comput Sci Inf Technol 2014;**5**(4).
Piao JT, Yan J. A network-aware virtual machine placement and migration approach in cloud computing. In: 9th International Conference on Grid and Cooperative Computing (GCC), 2010. IEEE.
Rasouli N, Meybodi M, Morshedlou H. Virtual machine placement in cloud systems using learning automata. In: 13th Iranian Conference on Fuzzy Systems (IFSC), 2013. IEEE.
Shu W, Wang W, Wang Y. Novel energy-efficient resource allocation algorithm based on immune clonal optimization for green cloud computing 2014. arXiv preprint arXiv:1405.4618A.
Srinivasan A, Quadir MA, Vijayakumar V. Era of cloud computing: a new insight to hybrid cloud. Procedia Comput Sci 2015;50:42–51.
Singh NA, Hemalatha M. Energy efficient virtual machine placement technique using banker algorithm in cloud data centre. In: International Conference on Advanced Computing and Communication Systems (ICACCS), 2013. IEEE.
Su K, et al. Affinity and conflict-aware placement of virtual machines in heterogeneous data centers.
Somani G, Khandelwal P, Phatnani K. Vupic: Virtual machine usage based placement in iaas cloud 2012. arXiv preprint arXiv:1212.0085.
Tang Z, et al. Dynamic forecast scheduling algorithm for virtual machine placement in cloud computing environment. J Supercomput 2014;**70**(3):1279–96.
Tordsson J, et al. Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers. Future Gener Comput Syst 2012;**28**(2):358–67.
Vu HT, Hwang S. A traffic and power-aware algorithm for virtual machine placement in cloud data center. Int J Grid Distrib Comput 2014;**7**(1):350–5.
Van HN, Tran FD, Menaud J.-M. Performance and power management for cloud infrastructures. In: IEEE 3rd international conference on Cloud Computing (CLOUD), 2010. IEEE.
Wang S-H, et al. EQVMP: energy-efficient and QoS-aware virtual machine placement for software defined datacenter networks. In: International Conference on Information Networking (ICOIN), 2014. IEEE.
Wang S, Gu H, Wu G. A new approach to multi-objective virtual machine placement in virtualized data center. In: Eighth international conference on Networking, Architecture and Storage (NAS), 2013 IEEE. IEEE.
Wang S, et al. Particle swarm optimization for energy-aware virtual machine placement optimization in virtualized data centers. In: International Conference on Parallel and Distributed Systems (ICPADS), 2013. IEEE.
Wu Y, Tang M, Fraser W. A simulated annealing algorithm for energy efficient virtual machine placement. In: IEEE international conference on Systems, Man, and Cybernetics (SMC), 2012. IEEE.
Xiaoli W, Zhanghui L. An energy-aware VMs placement algorithm in Cloud Computing environment. In: Second international conference on Intelligent System Design and Engineering Application (ISDEA), 2012. IEEE.
Xiao Z, et al. A solution of dynamic VMs placement problem for energy consumption optimization based on evolutionary game theory. J Syst Softw 2015;101:260–72.
Xu J, Fortes JA. Multi-objective virtual machine placement in virtualized data center environments. In: IEEE/ACM international conference on Green Computing and Communications (GreenCom), 2010 & international conference on Cyber, Physical and Social Computing (CPSCom). 2010. IEEE.
Yapicioglu T, Oktug S. A traffic-aware virtual machine placement method for cloud data centers. In: Proceedings of the 2013 IEEE/ACM 6th international conference on utility and cloud computing. 2013. IEEE Computer Society.
Yang T, Lee YC, Zomaya AY. Energy-efficient data center networks planning with virtual machine placement and traffic configuration. In: IEEE 6th international conference on Cloud Computing Technology and Science (CloudCom), 2014. IEEE.
Zhang Q, Cheng L, Boutaba R. Cloud computing: state-of-the-art and research challenges. J Internet Serv Appl 2010;1(1):7–18.
Zhang H, et al. Protecting private cloud located within public cloud. In: Global Communications Conference (GLOBECOM), 2013 IEEE. IEEE.
Zhang Y, Li Y, Zheng W. Automatic software deployment using user-level virtualization for cloud-computing. Future Gener Comput Syst 2013;29(1):323–9.
Zamanifar K, Nasri N, Nadimi-Shahraki M-H. Rate optimal virtual machine placement in cloud computing.
Zhang J, et al. SLA aware cost efficient virtual machines placement in cloud computing. In: IEEE International Performance Computing and Communications Conference (IPCCC), 2014. IEEE.
Zhuang Z, Guo C. OCPA: an algorithm for fast and effective virtual machine placement and assignment in large scale cloud environments. In: International conference on Cloud Computing and Big Data (CloudCom-Asia), 2013. IEEE.
Zheng Q, et al. Virtual machine consolidated placement based on multi-objective biogeography-based optimization. Future Gener Comput Syst 2015.
Zou T, et al. ClouDiA: a deployment advisor for public clouds. In: Proceedings of the VLDB endowment; 2012. VLDB endowment.
Zhang Q, Li M, Hu X. Network traffic-aware virtual machine placement with availability guarantees based on shadows. In: 14th IEEE/ACM international symposium on Cluster, Cloud and Grid Computing (CCGrid), 2014. IEEE.