An overview of the different methods for optimizing the virtual resources placement in the Cloud Computing.

Aristide Ndayikengurukiye¹, Abderrahmane EZ-Zahout¹ and Fouzia Omary¹
lemure.dede@gmail.com, abderrahmane.ezahout@um5.ac.ma and omary@fsr.ac.ma

E-mail: lemure.dede@gmail.com

Abstract. The emergence of Cloud Computing has driven users of new technologies to adopt it. This is made possible by the used technologies, namely the virtualization, which is the key element in Cloud Computing. The management and orchestration of virtual resources (the virtual machine placement and VNFs) in the Cloud remains a complex task, which, recently, attracted the interest of many researchers. This paper provides an overview of the techniques adopted to optimize the virtual machines placement and the virtual networks functions. Focused on different resources (CPU, memory, bandwidth and storage) applied in a virtualized environment. These techniques always target several approaches including the improvement of the QoS defined in the service level agreements (SLA), the good management of the energy used by physical resources, the allocation of resources and the management tasks in the data center.

1. Introduction
In recent years, many technologies have succeeded each other to alter the exploitation of IT resources. This has given rise to the Cloud Computing technology, which consists of providing virtualized services like information storage and calculations, sometimes via the net or internal network. Rather than acquisition high direct prices within the purchase of IT infrastructure and addressing computer code and hardware maintenance and upgrades, firms will source their IT has to the Cloud. This is often created potential by virtualization, which involves running multiple virtual machines connected to every different via virtual networks running on one physical machine. The implementation of those virtualized resources, as well as virtual machines, the square measure provided by computer code, referred to as hypervisors. In alternative words, virtualization could be a technique for sharing one physical instance of a resource or application across multiple purchasers. This technology not solely provides a virtual atmosphere for physical instances and applications, however additionally storage and networks. The diagram below summarizes the taxonomy of virtualization. We tend to show the various sorts of virtualization in keeping with every application domain and for every kind, we tend to assign the techniques that area unit applied thereto. On the far side of these techniques, it shows the requirement to contemplate the role of management and security in virtualization. The proliferation of Cloud Computing has led to the creation of large-scale data centers containing thousands of virtualized computing nodes and consuming huge amounts of electrical...
energy to run and cool physical resources. Several techniques are implemented through virtualization to properly manage virtualized resources and address the problems associated with these energy consumptions while ensuring a good quality of service in the datacenters. The VMs placement is done while solving the problem of server consolidation. To achieve this, Virtual Machine placement is the most widely used technique. It is subdivided into four steps including detection of host overload, detection of the host under load, selection, and migration of the VM [1].

With the evolution of technology, it is important to also highlight the placement of virtual network functions to establish the placement of virtual machines and maximize QoS optimization and energy exploitation. In 2012, The Europian Telecommunications Standards Institute established this new concept of virtualization of network functions and defined its basic architecture composed of three main elements: Virtual Network Functions (VNFs), VNF Infrastructure and MANO (Management and Orchestration). The VNF adds new functionality to all communication networks and requires a new set of management and orchestration functions. In existing networks, implementations of the virtual network function are often associated with the infrastructure on which they run. The NFV separates software implementations of network functions from the computing, storage, and network resources they use. Virtualization isolates network functions from these resources through a virtualization layer as shown in the figure below.

In this article, we provide an overview of the different methods for optimizing the virtual resources placement in the Cloud. The remainder of the paper is organized as follows. In Section II, we first present work on the placement of virtual machines. In section III, we present the different algorithms used in the virtual machines placement and the metrics used to evaluate them. In section IV, we present the different machine learning techniques used to predict resource usage in Cloud Computing. In Section V, we classified various VNF placement work into two broad categories. In Section VI, we summarize the work.
2. Related work

In the literature, several works related to the placement of virtual machines in the Cloud have been proposed. For example, Weijia S. et al. have proposed an approach that uses migration to dynamically allocate data center resources according to application needs while optimizing the number of servers used. They developed a bin-packing resource allocation algorithm that can both avoid overloads and apply IT[2].

In order to optimize the allocation of resources, N. Janani, R. D. Shiva Jegan and P. Prakash have implemented the multidimensional backpack algorithm to minimize the migration of virtual machines, thus maximizing resource utilization. To improve the efficiency of the backpack algorithm, they also used genetic algorithms to optimize the solution of the first algorithm[3].

Mosa and R. Sakellariou proposed a parameter-based VM placement solution that works in the range of workload and full reservation to mitigate problems on resource reservation and demand. This solution was implemented while considering multiple resources (CPU, RAM, and bandwidth) [4].

C. Ghribi, M. Hadji, and D. Zeghlache have implemented algorithms acting together to reduce energy consumption and migration costs with acceptable convergence time compared to the other algorithms being compared. The optimal allocation algorithm is solved as the bin packing problem to reduce energy consumption and is compared to the fit algorithm [5].

A. Abdelsamea, A. A. El-moursy, E. E. Hemayed, and H. Eldeeb proposed the use of hybrid factors to improve VM consolidation. They developed a multiple regression algorithm taking into account CPU, memory, and bandwidth for the detection of host overload. The proposed algorithm reduces power consumption while ensuring SLA compliance compared to the other regression algorithms being compared [6].

M. Alaul, H. Monil, and R. M. Rahman proposed an algorithm for the detection of host overloads based on the mean, median, and standard deviation of virtual machine usage. To improve the performance of selecting VMs to be migrated from one host to another, they proposed a method that makes the decision in an intelligent way[7].

M. R. Chowdhury and R. M. Rahman presented an approach to generate multiple copies of virtual machines without sacrificing QoS. An algorithm based on dynamic programming and local search was provided to determine the number of VM copies and place them on servers to minimize the total power cost in a Cloud Computing system. The proposed solution offers a
flexible method to increase the energy efficiency of the Cloud Computing system or even increase the availability of resources in data centers compared to existing solutions[8].

C. Wei, Z. H. Hu, and Y. G. Wang have proposed a heuristic algorithm to reduce energy consumption and optimize the resource waste problem. This algorithm reduces the number of physical machines in operation by maximizing and balancing the load that are applied. They have also introduced a penalty and reward mechanism to manage VMP of virtual machines on physical machines[9].

The table below summarizes the various works already discussed on the methods used in the placement of virtual machines. It shows the objectives of these various works, the resources considered (CPU, RAM, Storage and bandwidth), the methods used as well as the algorithms used for the comparison that were the subject of the previous works on virtual machine placement. It shows that several works did not consider all the resources, which creates a handicap during the VMP.

3. Virtual machine placement algorithms and their evaluation metrics

3.1. Evaluation algorithms

3.1.1. Bin packing Algorithm: The main objective of this algorithm is to insert a set of elements or items with a different size in a minimum number of boxes. The latter considers VMs as the objects that must be inserted in the physical machines considered as bins. With its variants such as First Fit, Best Fit, Worst Fit Decreasing, this technique is used to optimize the placement of virtual machines. The main objective of this algorithm is to reduce energy consumption and the allocation of virtual resources in data centers. The authors in[15, 16, 17, 18] used the different categories of the bin packing technique to map virtual machines on physical machines.

3.1.2. Stochastic Integer programming: It is a modeling and optimization method involving the notions of probability and which can be analyzed statistically but cannot be predicted accurately. It is useful in cases where the claims are not known, but the distribution of claims is known or can be estimated. Algorithms that use this technique can allocate resources in three phases: reservation, utilization and on demand. This approach has been applied in [2, 19, 20] for VM placement.

3.1.3. Genetic Algorithm: Generally a genetic algorithm is an optimization technique based on natural and genetic selection. In [21, 22, 23] the authors also employed this technique in the placement of virtual machines by making natural selection of the solution close to all possible solutions. These algorithms can also be considered as a bin Packing algorithm when it takes additional constraints while optimizing costs.

3.1.4. Constraint programming Algorithm: The problem of virtual machine placement in the Cloud can be formulated as a constraint programming problem by taking into account several constraints. It is a technique that allows to solve multi-objective problems in virtual machine placement [24, 25].

For the virtual machines placement, several constraints can be taken into account such as constrained SLA, constrained QoS, constrained capacity. The CPU, memory bandwidth and storage are the different dimensions considered for each physical machine. \(P_i\) the set of resources of each physical machine \(i\) composed of \(P_i^r = \{p_{i1}^r, p_{i2}^r, ..., p_{in}^r\}\) which represent the resources \(p_{ireg}, p_{iram}, p_{ibw}, p_{iro}\). VM allocation and resource consumption of resources used by VMs must be less than that of the physical machine:
| Authors                        | Objectives                     | Method used               | CPU | Memory | Storage | Bandwidth | Best than                            |
|-------------------------------|--------------------------------|---------------------------|-----|--------|---------|-----------|--------------------------------------|
| Weijia S. and al[2]           | Resource allocation and green computing | Algorithm bin packing     | ✓   | -      | -       | -         | Black box, Offline bin packing and VectorDot |
| Gao, Liang Liu [10]           | Waste of resources and energy consumption | Algorithm Ant colonies    | ✓   | ✓      | -       | -         | MMAS Multi objective genetic algorithm and binpacking |
| Janani and Shiva [3]          | Resource use, energy consumption | Genetic algorithm         | ✓   | ✓      | ✓       | ✓         | FF, FFD, Next made, Random Most fit and full |
| Mohammad A. and M. Rashedur [7]| Energy consumption and SLA     | fuzzy search algorithm     | ✓   | ✓      | -       | -         | IQR, RS, PBA MAD THR                  |
| Amany A. Ali A., E. and Hesham Elsayed E. [11]| Consumption of energy and SLA | multiple linear regression algorithm | ✓   | ✓      | -       | ✓         | single LRR regression algorithm and LR |
| Abdelkhalik Mosa and Rizos Sakellariou [4] | Resource utilization, SLA | Sorting algorithm          | ✓   | ✓      | -       | ✓         | -                                    |
| C. Ghribi, Mr. Hadji and D. Zeghlache [12]| Energy consumption | Constraint programming algorithm | ✓   | -      | -       | -         | Fit algorithm                         |

**Table 1.** Comparison of work on Virtual Machine Placement
3.2. Metrics for evaluating virtual machine placement algorithms

To evaluate the performance of the optimization algorithms used in the virtual machines placement, certain metrics are applied to them. In this part of the work we discuss some of these metrics.

**Energy consumption:** To evaluate the energy consumption in the data centers, the increase due to the use of the CPU must be taken into account. The relationship below shows that servers shut down when they are not in use, which reduces power consumption.

\[
P_j = \begin{cases} 
U_c j \times (P_{busy} j - P_{idle} j) + P_{idle} j, & U_c j > 0 \\
0, & \text{otherwise}
\end{cases}
\]  

**Service Level Agreement Violation:** A breach of the SLAV will occur when the cloud service provider fails to provide the service expected by customers. The SLAV metric can be calculated as follows:

\[
SLAV = SLATH \times PDM
\]  

**The SLATH:** This is the violation time for each active host. It is calculated as follows:

\[
SLATH = \frac{1}{N} \sum T_{oi} T_{ai}
\]

Where, \(N\) the number of hosts, \(T_{oi}\) is the total time host \(i\) used 100% of the CPU, which led to the violation SLAV, \(T_{ai}\) is the total duration of activity of host \(i\) for the machine virtual service[26].

The logic of SLATH is that virtual machines cannot get the required processor capacity due to 100% CPU usage of the active host.

**Performance degradation due to migration (PDM):** Live migration of virtual machines has a negative impact on the performance of the applications running on them. It is calculated as follows:

\[
PDM = \frac{1}{M} \sum \frac{C_{dj}}{C_{rj}}
\]

Where, \(M\) represents the number of VMs; In this relation \(C_{dj}\) is the is the estimate of the degradation of the virtual machine due to the migration of the VM\(_j\); And \(C_{rj}\) corresponds to the total CPU capacity requested by the VM\(_j\) during its lifetime.
4. Artificial intelligence for the optimization of virtual resources in the Cloud

In this part of the work, we analyze the different methods of artificial intelligence applied to the virtual machines placement. These can be directly applied to various virtual resource management techniques such as CPU usage, power consumption and migration. In [27], the authors have carried out a study of predictive techniques for virtual machine placement while specifying the different methods used. Other authors have shown the importance of these classification techniques for improving the overall efficiency and performance of data centers [28].

4.1. Artificial Neural network

ANNs are made up of layers of artificial neurons that process information using a transfer function and are interconnected on the basis of a specific network topology, each connection being associated with a weight and a bias. In the case of virtual machine placement, the average and maximum CPU usage, memory usage and other VM resources are defined as the input vectors. Each neural network with weighted connections receives these values from the hidden layer. As indicated in relation (9), the network will then calculate the sum of the weighted signals \( z_k \) for each neural network:

\[
    z_k = \sum_{j=1}^{m} w_{kj} x_j
\]

where \( x_1, x_2, ..., x_n \) are the input signals \( w_{k1}, w_{k2}, ..., w_{km} \) are the respective synaptic weights of the neuron \( k \); \( z_k \) is the output of the linear combiner due to the signal input. In our case the activation function is applied to the output signal of each neuron which transforms the range of the signal into a value between 0 and 1. The most commonly used activation function is a sigmoid function which is defined in the equation (10):

\[
    \gamma(z_k) = \frac{1}{1 + \exp(-a z_k)}
\]

where \( a \) is the slope parameter of the function, \( z_k \) is the activation value for the \( k \) neuron and \( \exp() \) is the exponential function. In [28], the authors show that for the case of classification problems, the network generates a probability whose input characteristics are between \( A \) and \( B \) when the signal is propagated in the network.

4.2. Support Vector Machines

SVM is a machine learning algorithm that is generally used for classification but it can also be used for regression. SVMs are the data points closest to the hyperplane. Data points in a data set which, if deleted, would change the position of the hyperplane in division and which in this case can be considered as preponderant elements of a data set. Like these other methods mentioned above, SVM improves the capacity for generalization by seeking to minimize the structural risk of the learning machine. The SVM is much more widely used when classifying small samples of data due to its fast training speed and high classification accuracy [29].

The SVM will separate the samples for a hyperplane for each binary problem. For the classifier to have a good classification accuracy, the SVM must obtain a suitable \( w \) and \( b \). The mathematical formula (11) shows how to model the non-linear decision function:

\[
    h(x) = sgn\left( \sum_{i=1}^{i=n} \alpha_i y_i K(X_i, X) + b \right)
\]
where \( k \) is a kernel function that is associated with the pair \((X_i, X)\). The kernel that is most used is the Gaussian kernel:

\[
K(X_i, X) = e^{-\frac{\|X_i, X\|^2}{2\sigma^2}}
\]  

(12)

### 4.3. K nearest neighbor (KNN)

K-NN is one of the methods used in machine learning. It is based exclusively on the choice of the classification metric, but it can also be used for regression metrics. To predict resource use, the closest neighbor \(k\) predicts the power used by each VM based on the similarity of characteristics and collects the training test mode [30]. To define the distance between the new value and the training value, the Euclidean distance is used, which is calculated as follows:

\[
d(x, y) = \sqrt{\sum_{j=1}^{n} (X_j - y_j)^2}
\]  

(13)

### 4.4. Tree decisions algorithm

Tree decisions algorithm is a method used in supervised learning based on the use of the decision tree as a predictive model. It is an algorithm that uses both classification tasks and regression. If the sample is homogeneous, the samples belong to the same class. The impurity measures this homogeneity of the sample. Two measurements are often used to measure impurity namely[31]:

**Entropy:** This refers to the amount of information needed to accurately describe certain samples. If the entropy is equal to 0 the sample is homogeneous which means that the elements are similar. And if the entropy is at the maximum so if it is equal to 1, it shows that all the probabilities \(p_i\) are equal. Mathematically, it is presented by the formula below:

\[
E(S) = -\sum_{i=1}^{n} p_i \times log(p_i)
\]  

(14)

where \(E(S)\) is the entropy of a data set, \(n\) represents the number of classes in the system and \(p_i\) represents the proportion of the number of instances that belong to class \(i\).

**Gini Index:** Is used instead of entropy to measure impurity. It measures inequality in the sample. If the Gini index is 0, it means that the sample is perfectly homogeneous, and if it is 1, it means that the inequality is maximum between the elements. Its mathematical formula is:

\[
I_G = 1 - \sum_{i=1}^{m} p_i^2
\]  

(15)

### 4.5. Linear regression

The linear regression algorithm is one of the algorithms of supervised learning. A variable, named \(X\), is considered as the predictive variable and a variable \(Y\) is considered as the target variable or the variable to be explained. For a simple linear regression, it is given by the following relation[32]:

\[
Y = aX + b + \epsilon
\]  

(16)

Where \(Y\) is the target variable, dependent random, \(a\) and \(b\) are coefficients to be estimated, \(X\) is the independent explanatory variable and \(\epsilon\) is a random variable that represents the error. Although the linear relationship is indeed present, the measured data do not exactly verify this relationship. In the mathematical model, we must take into account the observed errors noted \(\epsilon\).
which is the difference between the observed value $Y_i$ and the value $a_i X_i + b$ given by the linear relation.

4.6. Logistic regression
Logistic regression is a prediction algorithm used in classification. It is a linear classification model which is the counterpart of linear regression. The logistic function is defined by the relation [33]:

$$Y_i = \frac{1}{1 + e^{-k}}$$

where $y_i$ is the predicted classification for independent variables $x_1, x_2, \ldots, x_m$. The value $e$ is a constant and $y$ is defined as the summation of each independent variable multiplied by it is corresponding parameter [28].

5. Placement of Virtual Network Functions
In this part of the work, we first establish the link between Cloud Computing and NVFs before proposing various related works on the placement of the functions of virtual networks in a Cloud environment.

5.1. Relationship between VNF and Cloud Computing
VNF is the rapidly growing approach of performing network functions to servers precisely from virtual machines. Its main objective is to allow VNFs to be deployed and scaled up without putting new equipment in place. Its main objective is to allow VNFs to be deployed and scaled up without putting in place new equipment. In [34], the authors show that VNFs are not limited to telecommunications networks but rather extend to computer applications that can create large-scale communications services. Using Fig. 4, they compare Cloud Computing and VNF based on the different layers of service embedded in the Cloud and the architecture of virtual networks. It shows the correspondence of the IaaS layer of Cloud Computing to the NFVI part of the VNF architecture as well as the SaaS layer to the different VNFs.

Undeniably, VNFs cannot be dissociated from SDNs (Soft Define Networking), which are also emerging as an alternative to existing networks, by splitting the data plane and the control plane. These two approaches, although distinct, often remain associated and complementary.

---

| Layered Resources | Model | Example |
|-------------------|-------|---------|
| Business Applications, Web Services | SaaS | Facebook, Google Apps, Twitter, ZenDesk, Salesforce.com, Zoho Office |
| Software Framework | PaaS | Heroku, Azure, Google AppEngine, RedHat OpenShift, force.com |
| Virtual Machines | IaaS | OpenStack, Azure, Amazon Web Services (EC2, S3, DynamoDB), GoGrid, Rackspace |
| CPU, Storage, Bandwidth | | Data Centres |

**Figure 3.** Comparison of the architecture of NVF with the Cloud Computing service layers
In SDN networks, routers are relegated to simple packet switching devices, and one or more controllers manage the control plane. This simple approach allows network administrators to gain better control over their network traffic. Indeed, SDN makes it possible to (i) collect metrology data of nodes and links; (ii) centrally manage these nodes, allowing and facilitating network optimization compared to distributed control. Research shows that these are areas that are strongly linked through virtualization.

The problem of optimal placement of VNF is widely addressed in several research studies. Several strategies are proposed to solve the various associated problems. Despite this, several questions related to this optimal investment persist. Recent efforts have been made to better introduce VNF and have resulted in the classification of VNF placement into two broad categories including general network functions placement (subdivided into VNF forwarding graph, VNF chains placement and VNF replication) and specific network functions placement (subdivided into transcoder and cache, S-GW and P-GW, vDPI and firewall and CDN).

5.2. Study on the VNFs placement

5.2.1. VNF Forwarding Graph

VNF-FG is a complex structure that provides the logical connectivity between virtual or physical network function and providers service chaining (the process of steering traffic flows across VNFs).

In [35], the authors presented a new approach for joint placement of virtualized network functions (VNFs) and their associated chains over Cloud environments. Our Eigen decomposition based solution scales to thousands of nodes and links and in insensitive to request size and the number of requested VNF-FGs. The algorithm has a fairly flat execution time penalty dominated by the physical infrastructure (NFVI) size.

5.2.2. VNF Chains Placement

The VNF Chain Placement Problem (VNF-CPP) is another VNF placement-related problem, which is known to be NP-Hard. It is important to find placement schemes that can scale with the size of the problem and find good quality solutions. In [36], they formalize the network function placement and chaining problem and propose an Integer Linear Programming (ILP) model to solve it. Additionally, in order to cope with large infrastructures, we propose a heuristic procedure for efficiently guiding the ILP solver towards feasible, near-optimal solutions.

5.2.3. VNF Replication

In [37], they study the problem of VNF placement with replications, and especially the potential of VNFs replications to help load balance the network. In [38], the authors study the problem of VNF placement with replications, and especially the potential of VNFs replications to help load balance the network, while the server utilization is minimized. They present a Linear Programming (LP) model for the optimum placement of functions finding a trade-off between the minimization of two objectives: the link utilization and CPU resource usage. Carpio and Jukan study how to improve service reliability using jointly replications and migrations, considering the chaining problem inherent in NFV. While replications provide reliability, performing migrations to more reliable servers decreases the resource overhead. A Linear Programming (LP) model is presented to study the impact of active configurations on the network and server resources. Additionally, to provide a fast recovery from server failures, they consider N-to-N configurations in NFV networks and study its impact on server resources[38].

5.2.4. Transcoder and cachet

Nowadays, we find many streaming videos. The users challenge is to have a quick and non-interrupted access. In order to have a better QoS, the connection plays an important role.
Several elements can be taken into consideration to meet the challenge. The authors analyse some works studied on this subject. In [39] they laid out a network infrastructure that leveraged the storage and computing power of a cloud residing in the core for collecting network status and computing multi-path scalable video coding (SVC) streaming provisioning strategies. They show how OpenFlow is beneficial for optimizing the migration of transcoders. They show that not using this protocol would introduce an unacceptable interruption in the transmission of streaming[40].

5.2.5. S-GW and P-Gw Placement
Bagaa and al. have presented one of the important component of the carrier cloud vision, as to know Network Virtualization Function. They focused on the data anchor (PDN- GW) virtualization, and more specifically on how instantiating and assigning the virtual PDN-GW to UEs. Rather than using only geographical location, they proposed to consider applications/services type when selecting a PDN-GW to UEs[41]. In [42], they tackled the challenging problem of NF placement onto the cellular core. In this respect, we introduced a MILP and its relaxed variant for NF placement optimization, subject to capacity constraints, delay budgets between EPC components, and 3GPP-related restrictions. They further presented a greedy algorithm that strives to map NFs proximately to eNBs, inline with carriers common practice.

5.2.6. vDPI and Firewall
The authors formulate the vDPI placement problem as a cost minimization problem. They cast the problem as a multi-commodity flow problem. They then propose a centrality-based greedy algorithm and assess its validity by comparing it with the ILP optimal solution on random networks[43]. The authors tackle the VNFP problem with a focus on energy efficiency. Her work comprises developing an integer linear programming (ILP) formulation for the VNFP problem (subject to strict security and budgetary constraints) to minimize the energy consumption of servers, as well as implementation and evaluation of the model. They also showed that the security of the functions of virtual networks extends over several intrusion mechanisms[44].

5.2.7. CDN (Replication of Content Distribution Networks)
In [45], the study focused on virtual machine placement and flavors selection for different images in a CDNaaS platform. The platform manages a high number of videos deploying virtualized caches, transcoders, and streamers. A CDN slice owner can add videos specifying their resolutions and these videos are streamed to the end-users, consumers of the CDN slice. To create an efficient CDN slice, virtual machines hosting cache functions, transcoder functions, and streamer functions must be assigned adequate flavors and must instantiated at adequate locations. Similarly, the total cost to be paid by the CDN slice owner must be efficient and fair. They purpose, two solutions. The first one aims at minimizing the incurred total cost, whereas the second one aims at maximizing QoE. In [46], the authors have devised mechanisms for allocating an appropriate set of VNFs for each CDN slice to meet its performance requirements and minimize as much as possible the incurred cost in terms of allocated virtual resources. A mathematical model is developed to evaluate the performance of the proposed mechanisms. They first formulate the VNF placement problem as two Linear Integer problem models, aiming at minimizing the cost and maximizing the quality of experience (QoE) of the virtual streaming service. By applying the bargaining game theory, they ensure an optimal trade off solution between the cost efficiency and QoE.
6. Conclusion
In this article, we present an overview of the different methods of optimizing the investment of resources in the Cloud. First, we introduce the different works on virtual machine placement. Next, we presented the different algorithms used in the placement of virtual machines and the metrics applied to evaluate them. Thereafter, we presented different machine learning (ML) techniques used to predict the use of resources in virtual resource placement. Then, we classified the various VNF works into two main categories. In future work, we will implement these different algorithms applied in optimization. This will allow us to carry out a detailed comparative study in order to get the best out of them. In the same work, we will then propose and implement a new algorithm and compare it with the best algorithm we have found.

References
[1] A. Beloglazov and R. Buyya, Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers, Concurr. Comput. Pract. Exp., vol. 24(2012), pp. 1397–1420.
[2] W. Song, Z. Xiao, S. Member, Q. Chen, and H. Luo, Adaptive Resource Provisioning for the Cloud Using Online Bin Packing, vol. 63(2014), no. 11, pp. 2647–2660.
[3] Janani, N.; Shiva Jegan, R.D.; Dr. Prakash P, Optimization of Virtual Machine Placement in Cloud Environment Using Genetic, Research Journal of Applied Sciences, Engineering and Technology, Amrita Vishwa,” vol. 10(2015), pp. 274–287, 2015.
[4] A. Mosa and R. Sakellariou, Virtual Machine Consolidation for Cloud Data Centers Using Parameter-Based Adaptive Allocation, In Proceedings of the Fifth European Conference on the Engineering of Computer-Based Systems (ECBS ’17). Association for Computing Machinery, Article 16(2017), 1–10.
[5] C. Ghribi, M. Hadji, and D. Zeghlache, Energy efficient VM scheduling for cloud data centers: Exact allocation and migration algorithms, Proc. - 13th IEEE/ACM Int. Symp. Clust. Cloud, Grid Comput. CCGrid 2013, vol. (2013) pp. 671–678.
[6] A. Abdelsamea, A. A. El-Moursy, E. E. Hemayed, and H. Eldeeb, Virtual machine consolidation enhancement using hybrid regression algorithms, "Egypt. Informatics J., vol. 18(2017), pp. 161–170.
[7] M. Alaul, H. Monil, and R. M. Rahman, VM consolidation approach based on heuristics , fuzzy logic , and migration control,” J. Cloud Comput., 2016.
[8] M. R. Chowdhury, M. R. Mahmud, and R. M. Rahman, Implementation and performance analysis of various VM placement strategies in CloudSim, J. Cloud Comput., vol. 4(2015), pp. 1–21.
[9] C. Wei, Z. H. Hu, and Y. G. Wang, Exact algorithms for energy-efficient virtual machine placement in data centers, Futur. Gener. Comput. Syst., vol. 106(2020), pp. 77–91.
[10] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, Journal of Computer and System Sciences A multi-objective ant colony system algorithm for virtual machine placement in cloud computing,” J. Comput. Syst. Sci., vol. 79(2013), pp. 1230–1242.
[11] A. Abdelsamea, A. A. El-Moursy, E. E. Hemayed, and H. Eldeeb, Virtual machine consolidation enhancement using hybrid regression algorithms Virtual machine consolidation enhancement, "Egypt. Informatics J., vol. 18(2017), pp. 161–170.
[12] C. Ghribi, M. Hadji, and D. Zeghlache, Energy efficient VM scheduling for cloud data centers: Exact allocation and migration algorithms, Proc. - 13th IEEE/ACM Int. Symp. Clust. Cloud, Grid Comput. CCGrid 2013, vol. (2013) pp. 671–678.
[13] H. Goudarzi and M. Pedram, Energy-efficient virtual machine replication and placement in a cloud computing system, Proc. - 2012 IEEE 5th Int. Conf. Cloud Comput. CLOUD 2012, vol.(2012) pp. 750–757.
[14] A. C. Adamuthe and T. R. Nitave, Adaptive harmony search for optimizing constrained Resource Allocation problem, Int. J. Comput., vol. 17(2018), pp. 260–269.
[15] S. Jangiti, V. Vijayakumar, and V. Subramanivaswamy, Hybrid best-fit heuristic for energy efficient virtual machine placement in cloud data centers, EAI Endorsed Trans. Energy Web, vol. 7(2020), pp. 1–8.
[16] C. Lin, P. Liu and J. Wu, Energy-efficient Virtual Machine Provisioning Algorithms for Cloud Systems, 2011 Fourth IEEE International Conference on Utility and Cloud Computing, Victoria, NSW,vol. (2011), pp. 81-88.
[17] Z. Tang, Y. Mo, K. Li, and K. Li, Dynamic forecast scheduling algorithm for virtual machine placement in cloud computing environment, J. Supercomput., vol. 70(2014), pp. 1279–1296.
[18] S. Jangiti and S. Shiram, V.S., Scalable and direct vector bin-packing heuristic based on residual resource ratios for virtual machine placement in cloud data centers, Comput. Electr. Eng., vol. 68(2018), pp. 44–61.
[19] S. Chaisiri, B. Lee, and D. Niyato, *Optimal Virtual Machine Placement across Multiple Cloud Providers*, in Proceedings of IEEE Asia-Pacific services Computing Conference, vol.(2009), pp. 103-110.

[20] J. Zhou, Y. Zhang, L. Sun, S. Zhuang, C. Tang, and J. Sun, *Stochastic Virtual Machine Placement for Cloud Data Centers under Resource Requirement Variations*, IEEE Access, vol. 7(2019), pp. 174412-174424.

[21] A. S. Abohamama and E. Hamouda, *A hybrid energy-Aware virtual machine placement algorithm for cloud environments*, Expert Syst. Appl., vol. 150(2020), p. 113306.

[22] M. Tang and S. Pan, *A Hybrid Genetic Algorithm for the Energy-Efficient Virtual Machine Placement Problem in Data Centers*, Neural Process. Lett., vol. 41(2015), pp. 211–221.

[23] Mosa, A., Paton, N.W., *Optimizing virtual machine placement for energy and SLA in clouds using utility functions*, J Cloud Comp 5, 17 (2016). https://doi.org/10.1186/s13677-016-0067-7

[24] S. Kim and Y. ri Choi, *Constraint-aware VM placement in heterogeneous computing clusters*, Cluster Comput., vol. 23(2020), pp. 71–85.

[25] A. E. A. Elthraifi, S. H. Mohamed, and J. M. H. Elmirghani, *VM placement over WDM-TDM AWGR PON Based Data Centre Architecture*, arXiv e-prints, pages = arXiv:2005.03590,2020.

[26] P. D. Bharathi, P. Prakash, and M. V. K. Kiran, *Virtual machine placement strategies in cloud computing*, 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. (2017)i, pp. 1–7.

[27] Masdari, M., Zangakani, *M. Green Cloud Computing Using Proactive Virtual Machine Placement: Challenges and Issues*, J Grid Computing (2019). https://doi.org/10.1007/s10723-019-09489-9

[28] R. Shaw, E. Howley, and E. Barrett, *An intelligent ensemble learning approach for energy efficient and interference aware dynamic virtual machine consolidation*, Simul. Model. Pract. Theory, vol.(2019), p. 101992.

[29] G. Liu et al., *Predicting cervical hyperextension injury: A covariance guided sine cosine support vector machine*, IEEE Access, vol. 8(2020), pp. 46895–46908.

[30] T. Deepika and P. Prakash, *Power consumption prediction in cloud data center using machine learning*, Int. J. Electr. Comput. Eng., vol. 10(2020), pp. 1524–1532.

[31] D. M. Manias, M. Jammal, H. Hawilo, A. Shami, P. Heidari, and A. Larabi, *Machine Learning for Performance-Aware Virtual Network Function Placement*, IEEE International Conference on GLOBECOM,, Waikoloa, HI, USA, vol.(2019), pp. 1-6.

[32] L. Li, J. Dong, D. Zuo, and J. Wu, *SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Model*, IEEE Access, vol. 7(2019), pp. 9490–9500.

[33] Y. Jararweh, M. B. Issa, M. Daraghmeh, M. Al-Ayyoub, and M. A. Alsmirat, *Energy efficient dynamic resource management in cloud computing based on logistic regression model and median absolute deviation*, Sustain. Comput. Informatics Syst., vol. 19(2018), pp. 262–274.

[34] R. Mijuemi, J. Serrat, J. L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, *Network function virtualization: State-of-the-art and research challenges*, IEEE Cloud Computing, vol. 5, 17 (2016). https://doi.org/10.1109/s13677-016-0067-7

[35] M. Mechtri, C. Ghribi, and D. Zeghlache, *VNF placement and chaining in distributed cloud*, IEE Int. Conf. Cloud Comput. Cloud, pp. 376–383, 2017.

[36] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspar, *Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions*, Proc. 2015 IFIP/IEEE Int. Symp. Integr. Netw. Manag. IM 2015, vol.(2015)pp. 98–106.

[37] F. Carpio, S. Dihari, and A. Jukan, *VNF placement with replication for Load balancing in NFV networks*, 2017 IEEE International Conference on Communications (ICC), vol.(2017), pp. 1–6.

[38] F. Carpio, W. Bziuk, and A. Jukan, *Replication of Virtual Network Functions: Optimizing link utilization and resource costs*, 2017 40th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPTRO 2017 - Proc., vol.(2017)pp. 521–526.

[39] Z. Zhi, S. Li, and X. Chen, *Design QoS-aware multi-path provisioning strategies for efficient cloud-assisted SVC video streaming to heterogeneous clients*, IEEE Trans. Multimed., vol. 15(2013), pp. 758–768.

[40] P. Farrow, M. Reed, M. Glowiai, and J. Mambretti, *Transcoder Migration For Real Time Video Streaming Systems*, pp. 1–13, 2015.

[41] M. Baghaa, T. Taleb, and A. Ksentini, *Service-aware network function placement for efficient traffic handling in carrier cloud*, IEEE Wirel. Commun. Netw. Conf. WCNC, vol. 3(2014), pp. 2402–2407.

[42] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, *Network function placement on virtualized cellular cores*, 2017 9th Int. Conf. Commun. Syst. Networks, COMSNETS 2017, vol. (2017), pp. 259–266.

[43] M. Bouet, J. Leguay and V. CONAN, *Cost-based placement of vDPI functions in NFV infrastructures*, Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft), vol.(2015), pp. 1-9.

[44] I. P. Bolodurina and D. I. Parfenov, *Neural network model for optimize network work in the infrastructure of the virtual data center*,2017 25th Telecommun. Forum, TELFOR 2017 - Proc., vol. (2017), pp. 1–4, 2018.
[45] S. Retal, M. Bagaa, T. Taleb and H. Flinck, *Content delivery network slicing: QoE and cost awareness*, 2017 IEEE International Conference on Communications (ICC), Paris, vol.[1](2017), pp. 1-6.

[46] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck, *Optimal VNFs Placement in CDN Slicing over Multi-Cloud Environment*, IEEE J. Sel. Areas Commun., vol. 36(2018), pp. 616-627.