International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.1 (2017), pp.215-224
http://dx.doi.org/10.14257/ijmue.2017.12.1.18

Server Consolidation Using a Dynamic Model Approach

Nisha Chaurasia[1], Shashikala Tapaswi[1] and Joydip Dhar[1]

[1]*ABV-Indian Institute of Information Technology and Management, Morena Link Road, Gwalior, M.P., India*
{nisha, stapaswi, jdhar}@iiitm.ac.in

Abstract

The vigorous increase in the applicability of services through Cloud Computing has brought up major concern about management of a large number of servers supporting virtualization consuming high power. In respect of this, Server Consolidation approach leads to the reduction of these multiple numbers of servers into a very small count without any compromise in Quality of Service (QoS). Server consolidation manages the servers without degrading the services offered which needs to be revised timely in order to cope up with present high growth technological scenario. The paper considers the optimized server consolidation problem and proposes a Dynamic Server Allocation Problem (DSAP) model in contrast to Static Server Allocation Problem available in literature. The DSAP can afford the dynamic requests along with the ability to support parallelism.

Keywords*: Cloud Computing; Server Consolidation; Virtual Machine*

1. Introduction

The immense development in computation has limelighted the advanced form of computation as Cloud Computing. Cloud computing (or cloud for short) is a compelling technology [1-2]. In the morphological terms, cloud computing is the smart way of utilizing the computation and storage facilities at large scale on the on-demand basis where the user use the desired resources and/or storage elastically according to the requirement. According to National Institute of Standards and Technology (NIST), cloud computing is defined as a model for enabling convenient, on-demand network access to a shared pool of conFigureurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [3]. The services offered by cloud according to NIST are shown in Figure 1.

The success of cloud computing is by using Virtualization, where multiple instances of Virtual Machine (VM) run on the same physical hardware [4-6], allowing multiple tasks to be accomplished simultaneously. The involvement of virtualization enlightens the capabilities of Cloud Computing serving an ample number of user requests, performing large computations.

Large computation demands large number of service-offering servers and their maintenance. Moreover, management and deployment of these servers require high capital investment despite the presence of virtualization as backbone. Hence, provisioning of actively running servers is required such that every server at a workstation run to its maximum threshold leading to the reduction in the total count of active servers comparatively, without affecting the incoming requests. This brings into the concept of Server Consolidation in Cloud Computing. Server Consolidation thus can be understood as the management of under-utilized as well as over-utilized servers such that each server is utilized optimally. Consolidation improves the flexibility, agility, and adaptability to the fluctuating resource demands [7].

ISSN: 1975-0080 IJMUE
Copyright ⓒ 2017 SERSC

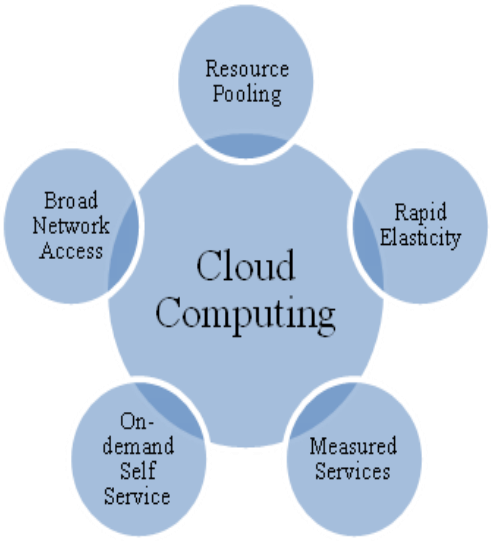

Figure 1. Cloud Services

Relating to the objective of consolidation and work done in consolidation, the paper proposes a novel mathematical model through which consolidation is expected to be achieved more effectively. The succeeding sections are as follows: Section 2 is a brief discussion about server consolidation, Section 3 throws light about the research done in consolidation so far, Section 4 is the explanation of the proposed model along with the assumption made, and Section 5 finally is the conclusion.

2. Server Consolidation

Server consolidation is an approach to the efficient usage of (physical) servers in order to reduce the total number of servers that an organization requires [8]. The approach can be well understood as the reduction in a total number of running servers intelligently such that no ongoing computation process gets affected. Consolidation means placing different jobs on the same physical machine that gets executed based on some scheduling policy [9-10]. It also defines reduction in the number of virtual servers running on a physical host. Consolidation can save management costs because it is much easier to manage a small number of machines than a large number of machines [7]. The main motive behind consolidation is to improve the utilization of each server aiding to effective computational cost and Quality of Service (QoS). There are two common aspects considered for consolidation viz. under-utilization and over-utilization. Under-utilized servers are those servers which lack computational workload and are continuously consuming resources resulting in server sprawl. Server sprawl is a situation in which multiple underutilized servers occupies more space and consume more resources that can be justified by their workload [1]. Over-utilized servers are those which are excessively loaded. As the average usages of many servers are low, then there is wastage of resources, and requires more staff to manage a large number of heterogeneous servers, thereby increases the total maintenance cost of the network [11]. Hence, every under-utilized and over-utilized server is made to share their workload such that all of the running servers reach to their threshold utilization provided no resource is left unused.

Server consolidation is a common method for energy optimization in cloud computing [12]. The server consolidation thus introduces the concept of VM Migration in cloud computing. The VM migration is roughly a load balancing criteria through which consolidation is ultimately achieved. As the live migration technology is widely used in the modern cloud computing data center, live migration of multiple virtual machines becomes more and more frequent [13]. The VM migration process involves selection of

 Copyright ⓒ 2017 SERSC

underutilized and/or over-utilized servers and thereafter sharing their workloads to servers capable of holding. Figure 2 shows the consolidation scenario.

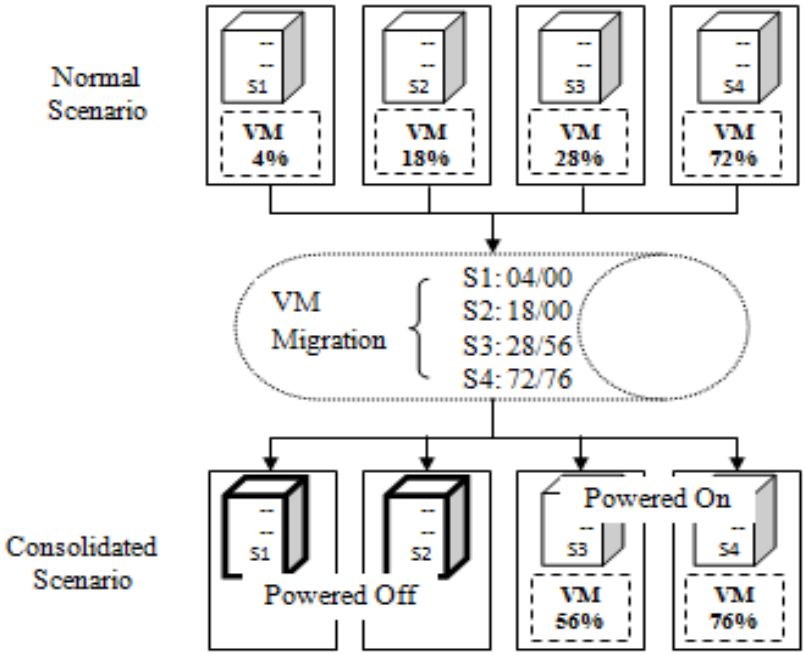

Figure 2. Server Consolidation Scenario in Cloud

3. Related Work

To reduce power consumption of data centers is an important issue, since cloud computing is becoming increasingly popular, and more data centers will be built [12]. The section explores the work performed so far to achieve an ideal level of server consolidation. Many magnificent types of research have been contributed confirming the consolidation attained to a remarkable value. The various profound consolidation schemes analyzed or proposed by the researchers may be categorized into three types:

3.1. Nature-inspired

In nature-inspired [4,14] the authors have performed a deep study on the behavior of various nature instances such as Swarm V-formation and Honeybee Hive formation. The authors have tried to learn the manner, these nature beings coordinate among each other such that the energy is conserved. Assuming the nature beings and their behavior to the Cloud scenario, the beings are considered as the servers or data centers and their behavior is imitated as the behavior which these Cloud servers would adopt to conserve energy. Hence, according to the way the nature beings coordinate their work, the data centers in Clouds are categorized into active servers, fully loaded active servers, idle servers, and power down servers. In the literature, it was found that managing the servers into these classes allow the easier assessment of how the energy can be conserved.

Copyright ⓒ 2017 SERSC

3.2. Parameter-based

In parameter-based [15, 11, 16, 12, 7, 17, 9, 18] the authors have focused on certain evaluation parameter viz. memory, CPU requirement. A parameter is pre-decided and based on this parameter the authors have proposed algorithms for the same. The purpose of this parameter is to determine the most appropriate server to which an incoming workload can be allotted. These parameters are formally based on NP-hard bin-packing problem such as best-fit, first-fit and heaviest-fit server finding.

3.3. Cost-based or Performance-based

In cost-based or performance-based [19, 13, 20-24] the authors consider a particular aspect while performing the computation in Cloud in order to minimize the cost of migration, CPU or disk usage or input/output events. They contributed to the consolidation with limiting these aspects, resulting in high performance leading to increased number of requests completed without violating the Service Level Agreement (SLA).

Although there are many significant contributions but still the need for improvement is encountered where higher consolidation is urged. The next section illustrates the mathematical model for expanding the consolidation to a greater level.

4. Proposed Dynamic Model

Keeping in view, the Static Server Allocation Problem (SSAP) discussed in [8], the paper introduces a Dynamic Server Allocation Problem (DSAP) with different level of servers. DSAP in contrast to SSAP considers real time/dynamic scenario for addressing the server allocation problem. The proposed model models the static scenario of SSAP to dynamic. The DSAP keeps control on the running servers according to their current capacities by categorizing the servers into low-end, medium-end, and high-end running servers. The model considers that there are k number of servers running and according to their current utilization and residual capacity they are considered to fall under j (from 1 to 3) type of servers. The service i among m servers is allotted to a server kj if the server is able to afford its workload and comprises of r resource(s) requested from the available R resources. The scenario is kept dynamic as there are changes in the functioning after every job computation. Hence, it is supposed that managing in this manner will provide optimized cost for server consolidation performed. The optimization of server consolidation problem is depicted through the schematic diagram, Figure 3.

 Copyright ⓒ 2017 SERSC

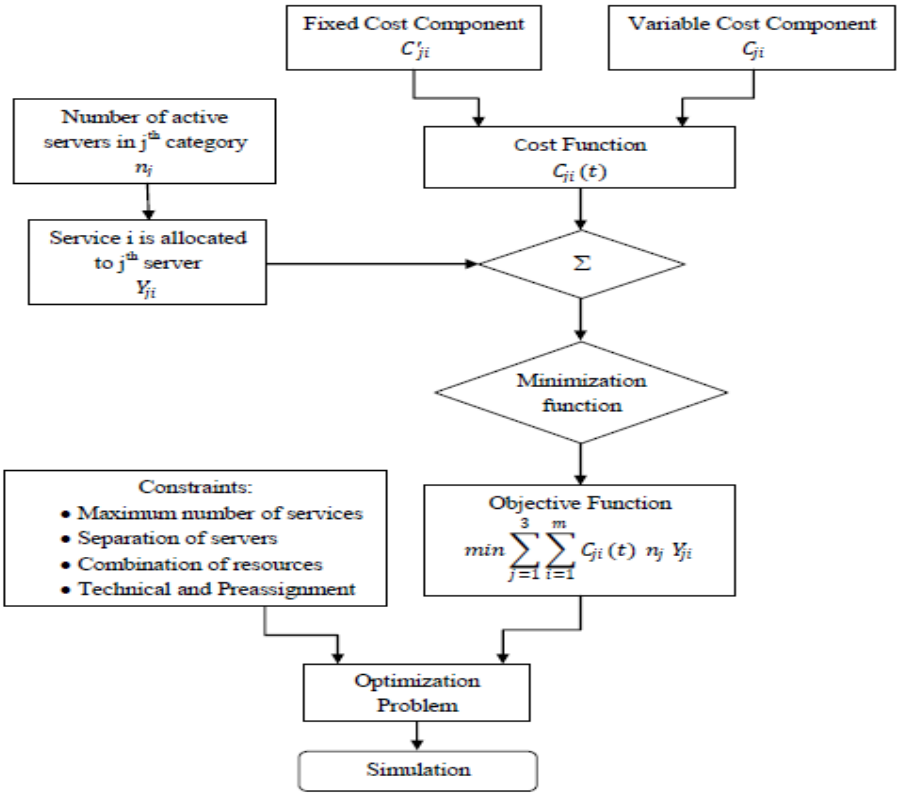

Figure 3. Schematic Diagram

In view to the schematic flow and to make more realistic cost optimization problem, DSAP considers following assumptions:

1. There are three levels of servers with respect to their computational ability.
2. The running cost of the server is defined in terms of fixed and variable components at each server level.

$$C_{ji}(t) = C_{ji_0} + C_{ji}.t, \tag{1}$$

where, C_{ji} is the cost function with fixed component $C_{ji0}$and variable component C_{ji}.t, and t is the time of utilization of a server.

3. The cost is dependent on the time t for which a server gets utilized.
4. The jobs execution in the model is expected to support parallel task performance.

On the basis of these assumptions and considering the dynamic environment, the objective function for DSAP is formulated as:

$$\min \sum_{j=1}^{3} \sum_{i=1}^{m} C_{ji}(t)\ n_j\ Y_{ji}, \tag{2}$$

where, C_{ji} is the cost, n_j is the number of servers in j^{th} category (n_j varies from 0 to kj), and Y_{ji} is the indication that server j is assigned with service i.

The function attempts to minimize the total cost of allocation problem, such that that each request is allocated required resources and no active server is left unallocated. With respect to this, the objective function is followed such that

$$\sum_{k=1}^{l} Y_{ji} = 1, \qquad \forall i \in m, \tag{3}$$

$$\sum_{i=1}^{m} u_{ir} Y_{ji} \leq s_{jr}\, n_j, \qquad \forall i \in m,\ \forall r \in, \tag{4}$$

Copyright ⓒ 2017 SERSC

where, u_{ir} defines the units of resources required by a service and s_{jr} defines the resource capacity of the server and

$$n_j, Y_{ji} \in \{0,1\}, \qquad \forall i \text{ and } j.$$

It is ensured that objective function aims at minimizing the server costs such that each service is assigned only once for execution and the server capacity does not exceed because of multiple workloads from multiple jobs.

Motivated from the work in [8], the DSAP can be extended by imposing certain constraint so that cost optimization is attained effectively. These constraints are:

1. Maximum number of service constraint: It is expected that maximum number (m_k) of services allocated to a server should be restrictive such that administrative time, effort in server failure event, etc are limited.

$$\sum_i Y_{ji} \leq m_k. \tag{5}$$

2. Separation constraint: There may be few services which need to be allocated to different servers for the purpose of security or for some technical reasons.

$$\sum_{i \in S} Y_{ji} \leq 1, \tag{6}$$

where, S is a subset of services allotted to servers following separation constraint.

3. Combination constraint: It is preferred that all the services (denoted by e) in subset S should be allotted to the same server as there is increased inter-application communication or the operating system requirements remain same thereafter also.

$$-(|S| - 1).Y_{je} + \sum_{S-\{e\}} Y_{ji} = 0, \qquad \forall e \in S. \tag{7}$$

4. Technical constraint and Pre assignment constraint: There may be service which requires particular server attribute (represented by a subset of servers, K pertaining this attribute) hence it is recommended that only one server from R is allotted to the service. Also, if $|K| =$, then the case is supposed to be of pre assignment constraint.

$$\sum_{k \in K} Y_{ji} = 1. \tag{8}$$

5. Experimentation and Results

The experimentation has been performed in cloud scenario with 50 servers running. It is assumed that the servers are capable of addressing the incoming tasks. It is also made sure that each task is allocated to only one server such that no server is left idle. The servers are made to run in idle state initially to estimate the initial cost and thereafter are allotted tasks adding some additional cost. The analysis has been performed for two cases.

Case1 offers random allocation of tasks where the servers with their initial running cost are made to execute the incoming tasks in First Come First Serve (FCFS) manner. This leads to the cases where less initial cost servers are allotted task with less cost or vice versa. Here, it is to be noticed that this strategy of allocation can lead to under-utilization as well as over-utilization conditions. Hence, Case2 based on the proposed work restricts the under/over utilization conditions. The tasks allocation with categorization of servers (high-end, medium-end, low-end) makes the servers address a task with minimum total cost.

Figure 4 and Figure 5 show the variation in the cost of servers at initial state with the cost of execution of task for both cases.

 Copyright ⓒ 2017 SERSC

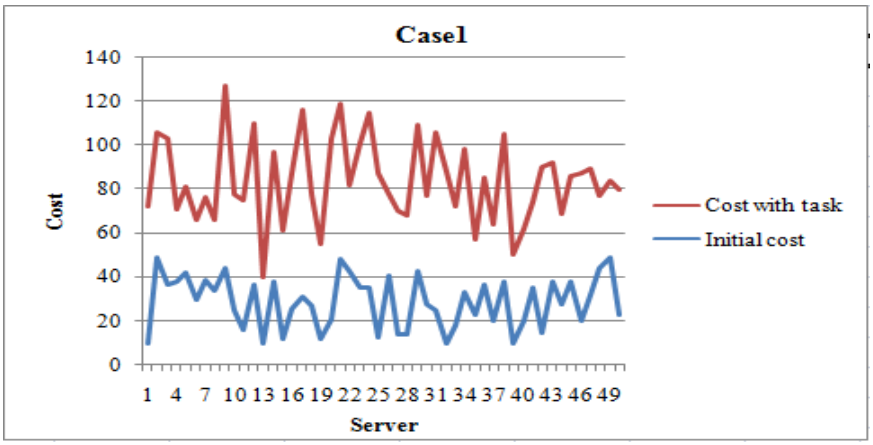

Figure 4. Cost Comparison for Case1

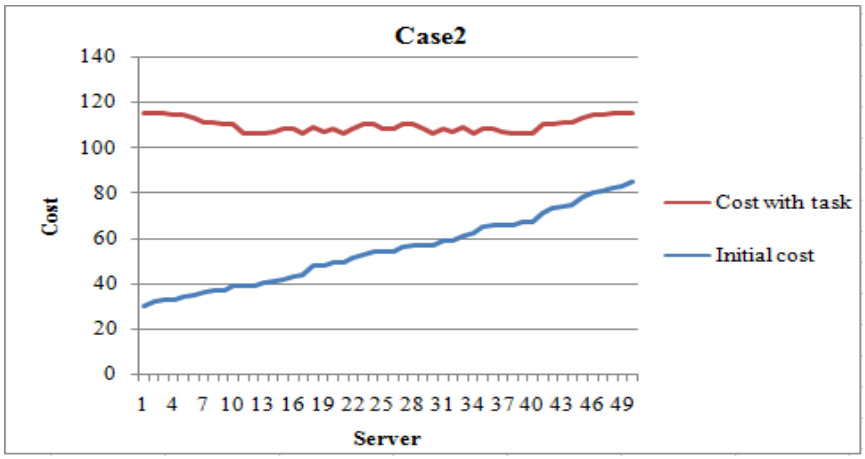

Figure 5. Cost comparison for Case2

Figure 6 shows the comparison of the total cost for the two cases. It is observed that the total cost for Case1 is higher in comparison to Case2. Keeping the servers run task with less cost makes the server consume less execution cost. Hence Case2 excels Case1 where total cost for Case2 is more linear than Case1.

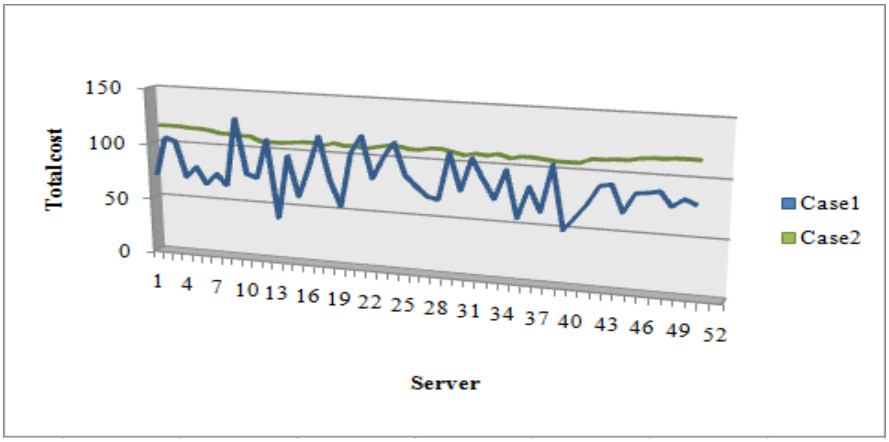

Figure 6. Cost Comparison for Case 1 and Case2

6. Conclusion

Copyright ⓒ 2017 SERSC

Following the work of static cost optimization, the paper proposes the new dynamic cost optimization problem, DSAP for server consolidation. The problem analyzing the realistic scenario defines the objective function for server consolidation. The categorization of servers according to their workload ability is assumed to provide faster handling of computation requests. It is expected that the different constraints and dynamic behavior offers a higher cost optimization approach in cloud computing. In addition, the dynamic environment ensures the possibility of affording parallelism. The experimental analysis proves the optimization by providing less cost of execution for active servers.

References

[1] S. Thakur, A. Kalia and J. Thakur, "Performance Evaluation of Server Consolidation Algorithms in Virtualized Cloud Environment with Constant Load", Int. Journal Advanced Research in Computer Science and Software Engineering, vol. 4, no. 3, **(2014)**, pp. 555-562.
[2] D. Shen, J. Luo, F. Dong, X. Fei, W. Wang, G. Jin and W. Li, "Stochastic Modeling of Dynamic Right-sizing for Energy-Efficiency in Cloud Data Centers", Future Generation Computer Systems, Elsevier, vol. 82, **(2015)**, pp. 82-95
[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing", National Institute of Standards and Technology (Article), **(2009)**.
[4] A. N. Singh and M. Hemalatha, "Cluster based Bee Algorithm for Virtual Machine Placement in Cloud Data Center", Journal Theoretical and Applied Information Technology, vol. 57, no. 3, **(2013)**, pp. 1-10.
[5] S. Esfandiarpoor, A. Pahlavan and M. Goudarzi, "Structure- Aware Online Virtual Machine Consolidation for Datacenter Energy Improvement in Cloud Computing", Computers and Electrical Engineering, Elsevier, vol. 42, **(2015)**, pp. 74-89.
[6] R. W. Ahmad, A. Gani, S. H. A. Hamid, M. Shiraz, A. Yousafzai and F. Xia, "A Survey on Virtual Machine Migration and Server Consolidation Frameworks for Cloud Data Centers", Journal of Network and Computer Applications, Elsevier, vol. 52, **(2015)**, pp. 11-25
[7] S. Lee and S. Sahu, "Efficient Server Consolidation considering Intra-Cluster Traffic: In: Global Telecommunications Conference (GLOBECOM 2011)", IEEE, **(2011)**, pp. 1-6.
[8] B. Speitkamp and M. Bichler, "A Mathematical Programming Approach for Server Consolidation Problem in Virtualized Data Centers", Transactions on Services Computing, IEEE, vol. 3, no. 4, **(2010)**, pp. 266-278.
[9] S. Nevithitha and V. S. S. S. Sriram, "Consolidated Batch and Transactional Workloads using Dependency Structure Prioritization", Int. Journal Engineering and Technology (IJET), vol. 5, no. 2, **(2013)**, pp. 1328-1334.
[10] K. S. Rao and P. Thilagam, "Heuristics based Server Consolidation with Residual Resource", Future Generation Computer Systems, Elsevier, vol. 50, **(2015)**, pp. 87-98.
[11] A. R. Abdulgafer, P. N. Marimuthu and S. J. Habib, "Network Redesign through Servers Consolidations", In: Information Integration and Web based Applications and Services (iiWAS2009), **(2009)**, pp. 623-627.
[12] Y. Ho, P. Liu and J. Wu, "Server Consolidation Algorithms with Bounded Migration Cost and Performance Guarantees in Cloud Computing", In: Fourth International Conference on Utility and Cloud Computing, IEEE, **(2011)**, pp. 154-161.
[13] K. Ye, X. Jaing, D. Huang, J. Chen and B. Wang, "Live Migration of Multiple Virtual Machines with Resource Reservation in Cloud Computing Environments", In: International Conference on Cloud Computing, IEEE, **(2011)**, pp. 267-274.
[14] C. B. Pop, I. Anghel, T. Cioara, I. Solemie and I. Vartic, "A Swarm-inspired Data Center Consolidation Methodology", In: 2nd International Conference on Web Intelligence, Mining and Semantics, ACM, vol. 41, **(2012)**.
[15] G. Khanna, K. Beaty, G. Kar and A. Kochut, "Application Performance Management in Virtualized Server Environments: In: Network Operations and Management Symposium (NOMS)", IEEE/IFIP, **(2006)**, pp. 373-381.
[16] Z. Gong and X. Gu, "PAC: Pattern-driven Application Consolidation for Efficient Cloud Computing. In: International Symposium on Modeling", Analysis and Simulation of Computer and Telecommunication Systems, IEEE/ACM, **(2010)**, pp. 24-33.
[17] C. Mastroianni, M. Meo and G. Papuzzo, "Probabilistic Consolidation of Virtual Machines in Self-Organizing Cloud Data Centers", Transactions on Cloud Computing, IEEE, vol. 1, no. 2, **(2013)**, pp. 215-228.
[18] A. Corradi, M. Fanelli and L. Foschini, "VM Consolidation: A Real Case based on OpenStack Cloud", Journal of Future Generation Computer Systems, Elsevier, vol. 32, **(2014)**, pp. 118-127.
[19] S. Srikantaiah, A. Kansal and F. Zhao, "Energy Aware Consolidation for Cloud Computing", In: Power aware computing and systems (HotPower'08), ACM, **(2008)**, pp. 1-10.

 Copyright ⓒ 2017 SERSC

[20] M. Maezolla, O. Babaoglu and F. Panzieri, "Server Consolidation in Clouds through Gossiping", In: International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), IEEE, **(2011)**, pp. 1-6.
[21] Z. Huang, D. H. K. Tsang and J. She, "A Virtual Machine Consolidation Framework for MapReduce enabled Computing Clouds", In: International Teletraffic Congress (ITC), IEEE, vol. 24, **(2012)**, pp. 1-8.
[22] X. Liu, C. Wang, B. B. Zhou, J. Chen, T. Yang and A. Y. Zomaya, "Priority-Based Consolidation of Parallel Workload in the Cloud", Transactions on Parallel and Distributed Systems, IEEE, vol. 24, vol. 9, **(2013)**, pp. 1874-1883.
[23] A. Beloglazov and R. Buyya, "Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of service Constraints", Transactions on Parallel and Distributed System, IEEE, vol. 24, no. 7, **(2013)**, pp. 1366-1379.
[24] Y. Xia, M. C. Zhou, X. Luo, Q. Zhu, J. Li and Y. Huang, "Stochastic Modeling and Quality Evaluation of Infrastructure-as-a-Service Clouds", Transactions on Automation and Engineering, IEEE, vol. 1, no. 99, **(2013)**, pp. 1-9.

Copyright ⓒ 2017 SERSC

Copyright ⓒ 2017 SERSC