Framework of Resource Management using Server Consolidation to Minimize Live Migration and Load Balancing

Alexander Ngenzi[*]
[*]Research Scholar
Computer Science Engineering, Jain University, Bangalore, India

Selvarani R[**]
[**]Professor and Head
Dept. Computer Science Engineering, Alliance University, Bangalore, India

Suchithra R[***]
[***]Professor and Head: Dept.Master of Science in Information Technology,
Jain University, Bangalore, India

***Abstract*—Live Migration is one of the essential operations that require more attention to addressing its high variability problems with virtual machines. We review the existing techniques of resource management to find that there are less modeling to solve this problem. The present paper introduces a novel framework that mainly targets to achieve a computational effective resource management technique. The technique uses the stochastic approach in modelling to design a new traffic management scheme that considers multiple traffic possibilities over VMs along with its switching states. Supported by an analytical modelling approach, the proposed technique offers an efficient placement of virtual machine to the physical server, performs the computation of blocks, and explores reduced resource usage. The study outcome was found to possess potential reduction in live migration, more extent of VM mapping with physical servers, and increased level of capacity.**

Keywords—Resource Management; Live Migration; Virtual Machine; Load Balancing; Cloud Computing

I. INTRODUCTION

The introduction of the cloud computing offers a change in the process of accessing as well as retrieving the data from multiple sources of clusters spread over a large geographic area [1]. The process of virtualization has played a significant contributory role for offering both data and service availability [2]. All Virtual Machines (VMs) operate in a highly integrated process resulting in an improvement in resource utilization to cater up to the massive demands on online users [3]. Owing to expensive-characteristics of cloud-based resources, it is quite imperative to perform optimization of the resource using server consolidation. In this regard, the placement of the VM is quite important to be considered when it comes to server consolidation [4] as the inappropriate placement of VM will result in maximum drainage of resources. There are various studies e.g. [5][6] that has discussed the variability problems of the traffic are existing over VM. The prime reason for this variability is the increasing adoption of enterprises with many programs that demand consistent and reliable performance. The spiky traffic over VM will represent maximized variability that assists in implicating statistical processes to evaluate the utilization trends [7][8]. A closer look into the existing techniques shows that scheduling of usual traffic utilizes the elasticity properties of cloud, but it is necessary to meet positive dynamic demands of resources to avoid overheads [9]. Therefore, live migration policies [10] and local resizing [11] are the frequently used techniques for catering up the dynamic demands of peak traffic condition. The configuration of the VM is adaptively changed in local resizing process whereas live migration results in placing some VMs to those physical servers that are found to be idle for a certain period. Although, live migration of VM is one of the most important processes associated with VM to provide seamless service delivery, it is carried out at the cost of higher resource utilization that finally results in potential downtime of some important services offered by the associated VM. Therefore, there is a need of carrying out an investigation to explore the best possibility of resource management by evolving up with a robust solution to living migration problems in the cloud along with load balancing. Therefore, the present paper has introduced one such technique which applies an analytical modeling to maintain a better level of equilibrium between live migration and efficient resource management as well as with better load balancing to the incoming traffic. The paper is arranged as per: Section 1.1 discusses the background of the study, Section 1.2 discusses the problem identified in the study, and Section 1.3 presents a brief discussion of proposed system. Section II discusses the algorithm implementation followed by analysis of result accomplished from the study in Section III. Finally, Section IV makes some concluding remarks.

A. Background

Study towards efficient resource management over cloud environment is not a new, and there has been the various amount of work has been already carried out till date. However, we will update only the most recently explored literature published in last 5 years about resource management, live migration, and server consolidation in this section. Zhu et al. [12] have incorporated a software-engineering based technique to perform scheduling of resources. Kumar and Saxena [13] have presented a study on quantitative analysis about the migration of VM along with its associated factors. Saraswathi et al. [14] have developed a technique of resource allocation to perform selection and execution of high priority task. The review paper was agreed using time and numbers of processing elements and host number. Wood et al. [15] have presented a model of live migration using dynamic pooling mechanism. The study has also presented an optimization

principle to reduce the storage cost as well as the memory of VM. Panda et al. [16] have discussed an algorithm that targets multiple environments of cloud based on the smoothening concept. The evaluation of the study was carried out using a bigger dataset of heterogeneous types. Study towards live migration problem has been carried out by Song et al. [17] where the authors have emphasized on forwarding the memory pages to retain cost effectivity in channel capacity as well as to reduce the total time of migration. Selvarani and Sadhasivam [18] have presented a task scheduling scheme over the cloud to perform mapping of the required resources. The cost of resources, as well as performance of computation, is estimated by the presented technique, and its outcome was analyzed with respect to time and cost. Nahir and Order [19] have introduced a formal framework for load-balancing using unique management policies of VM. The study outcome was testified using overhead on the mean queue. Kao et al. [20] have introduced an involuntary decision-making technique for facilitating the better process of live migration. Taking the case study of private cloud, the authors have implemented it as an experimental prototype. The study outcome was explored with better scalability; energy saving features as well as load balancing characteristics. Wei et al. [21] have addressed the problem of resisting utilization of skewed resources over physical server using resource-based prediction approach. The study has also presented a completely new technique of resource allocation of heterogeneous types for catering up multiple demands on the cloud-based networks. Yue and Chen [22] have presented a non-probabilistic technique to address the problems of VM placement over the data centers. The study outcome has shown energy efficiency as well as leads to minimization of physical servers to approximately 20%. Caton et al. [23] have used an open-source framework that uses the potential networking attributes of the social network to carry out resource allocation in the cloud. The presented study also uses stochastic modelling of node participation process. Assessment of the server consolidation was carried out by Chang et al. [24] where the problem of selection of a precise hypervisor is discussed for specific virtualization of the server. Study on dynamic allocation of resources was also carried out by Yang et al. [25] to perform autonomous migration of the jobs among the VMs depending on the amount of load. The result was assessed using time with increasing size of problem and CPU utilization using OpenNebula. Perumal and Murugaiyan [26] have adopted an optimization technique to address the problems of VM placement and consolidation of the server. Eramo et al. [27] have presented a unique architecture to solve the problem of dimensioning of server resources using optimization technique. The study outcome was found to possess better energy saving features. Study towards live migration problem was discussed by Sarker and Tang [28] has proposed an effective scheduling of VM migration policies. Ye et al. [29] have presented a framework using profiler for the purpose of minimizing physical servers along with retention of better performance of different traffic.

B. The Problem

From the previous section, it can be seen that there are various techniques towards problems related to resource allocation, live migration, server consolidation. The technical pitfalls in majority of the approaches are as follows:

- The majority of the study is focused on increasing live migration as means of server consolidation. Unfortunately, an increase of live migration also results in performance degradation while working under constraints, which is still not addressed in the existing system.
- Only a few studies have used the potential feature of stochastic and probability theory in modeling, which could be used for better visualization of different dynamicity of the traffic. In short, traffic modeling is found not to be emphasized much.
- There is lesser extent of modeling relationship between usage of physical servers, live migration, and capacity
- The proposed study identifies the above-mentioned problems and considers to solve it by its presenting analytical modeling approach. The next section briefs about the proposed solution adopted to counter-measure the identified problems.

C. The Proposed Solution

The purpose of the proposed system is to introduce a novel framework that can perform an effective server consolidation with retention of minimized live migration of VM and increased load balancing system over data centers in the cloud environment. The present work is a continuation of our prior work being carried out [25]. The complete implementation of the proposed system follows analytical modelling approach. Fig.1 highlights the proposed scheme of [FRMS].

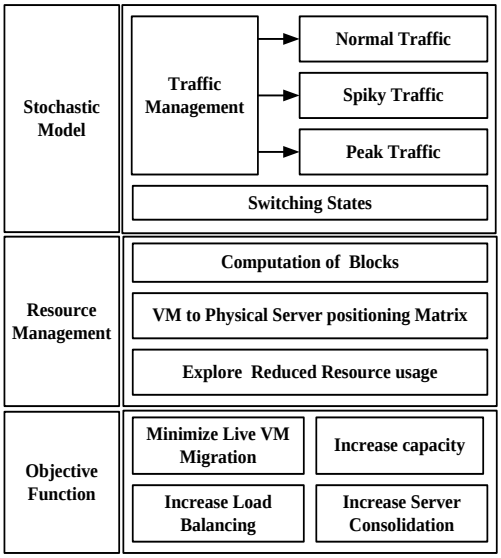

Fig. 1. Schematic Diagram of Proposed System

The proposed system introduces an empirical modeling using stochastic approach applicable for *traffic management* and *switching state* designing. The study formulates three different states of traffic i.e. *normal traffic*, *spiky traffic*, and *peak traffic* situation modeled using stochastic approach. The study also uses a probability parameter to represent its switching states i.e. states of ON and OFF corresponding to higher and normal traffic situation respectively. The resource management block mainly consists of i) *computation of blocks*

ii) *VM to physical server positioning matrix*, and iii) *exploring reduced resource usage*. The block will represent an effective serving window to perform load balancing by minimizing its number. Computation of blocks is carried out considering both the switching states with cut-off capacity value. The computed blocks assist in finding the less number of reserved spaces for physical servers. The positioning matrix assists in the allocation of VM to the respective physical serves by the number of VMs, the specification of physical servers, switching states, and capacity factor. Finally, exploration of minimal resource usage is carried out by developing a new matrix that can record only the minimal blocks needed to be allocated by the physical server on a defined spike of workload. The entire evaluation of the server consolidation is carried out by comparing mainly normal traffic and spiky traffic. Finally, the objective function is developed that is responsible for exploring mapping of VM to the physical server to minimize an event of live migration over dynamic and unpredictable traffic over cloud environment. The prime goal of the proposed approach is to ensure the existence of space with approximately zero waiting time over the load balancing system (i.e. queue). In this process, each VM that possesses its individual blocks will also be subjected to be reduced as minimal number as possible while maintaining the constraints of performance satisfied. Hence, the objective function balances minimization of live VM migration, increases capacity, maximizes load balancing system, and finally enhances server consolidation. The next section discusses algorithm implementation.

II. Algorithm Implementation

The prime purpose of the proposed algorithm is to ensure an effective resource management to be taking place in the cloud data centers with a core goal of accomplishing server consolidation. The proposed algorithm takes the input of T_n (Normal Traffic), T_v (Variable Traffic), T_h (High Traffic), η (Samples), α (number of physical servers), S_1 (Switching state-1(off→on)), S_2 (Switching state-2(on→off)), *mr* (Minimum resources), *d* (Highest number of VM permissible for physical servers), φ (capacity of host machine), τ (Capacity overflow), ρ (number of partition), N_{mig} (Number of Migration). The algorithm after processing results in live migration (denoted by N_{mig}++). The steps of the algorithms are as follows:

Algorithm for incentive allocation

Input: T_n, T_v, T_h, η, α, S_1, S_2, *mr*, *d*, φ, τ, ρ, N_{mig}

Output: Live Migration (N_{mig}++)

Start

1. init T_n, T_v, T_h,
2. $w_{load} = T_{n1} + (T_{n2} - T_{n1})$*arb(1, η);
3. for i=1:k
4. for j=1: α
5. for r=0:j
6. $op = S_1^{r(1-S_1)^{(j-r)}} . S_2^{(j-i+r).(1-S_1)^{(k-j-r)}}$
7. end
8. end
9. end
10. mr←1-[(v_1+v_2)/2]
11. for k=1:d
12. min_res←minimum_resource_block (k, S_1, S_2, mr)
13. end
14. sort(min_res)
15. for i = 1:length(PS)-1; //PS→sort(min_res)
16. X(i) = max([T_v(i),max(T_v)])*min_res(i+1)+T_n(i)+∑(T_n<φ);
17. end
18. $w_{load=}$ w_{load+1}
19. for g=2:G
20. if k ≠g
21. ρ(k)= ρ(k-1)+ ρ(k+g)
22. end
23. r = min_res(xi)*max(T_h(min(j,G)))
24. if r<r_{min}
25. r_{min}←r
26. else, N_{mig}++
27. End

End

The algorithm starts by empirically generating the traffic (Line-2). The complete algorithm performs three types of conditional checks i.e. i) of T_n=T_v, ii) T_n>T_v, and iii) T_n<T_v. Using state-based transition probability, the algorithm determines a probability factor *op* for assessing an effective load balancing (Line-6). To overcome server consolidation problem, the algorithm computes a minimum number of block

mr (Line-10), which are obtained from v1 and v2 that corresponds to the sum of all stationary distribution from 1 to *(k-1)* and 1 to *k* respectively. The stationary distribution is obtained by applying row reduction method [30]. Minimum resource block is then computed considering the input parameters of i) *k* (all numbers of VMs allowed on physical servers), ii) switching state S_1 from OFF state to ON state, iii) switching state S_2 from ON state to OFF state, and minimum resource *mr* (Line-12) that is finally sorted to obtain the better resource block (Line-14). The constraint of real Virtual Machine (VM) migration with an efficient load balancing is addressed empirically by computing X for all the physical servers (Line-16). It will mean allocation of a specific VM on the initial physical server in case the entire value obtained by X is found less than the capacity of host machine φ (Line-16) followed by incrementing traffic (Line-18). Hence, the summation of X leads to estimating a total number of used physical servers. Finally, live migration is optimized as follows viz. G is computed that represents the size of T_v is estimated (Line-19), if the number of VM permissible for physical servers (k) is not same as variable g than the algorithm empirically generates all g partitions (Line-21). Finally, minimum resource (r_{min}) is computed (Line-23) and conditionally checked to perform live migration of the VMs (Line24). Hence, as an output, it computes some live migration required to perform server consolidation.

III. Result Analysis

This section discusses the results obtained from the proposed study. The study outcome is evaluated in three different conditions of normal traffic, existing traffic, and proposed the system. The assessment was carried out for observing some used physical machines (or servers), capacity overflow, some migration, and processing time. The existing system of traffic management performs reservation of the certain specific proportion of resources on each physical server that can be considered to be permissible server consolidation strategy without any apriori of traffic.

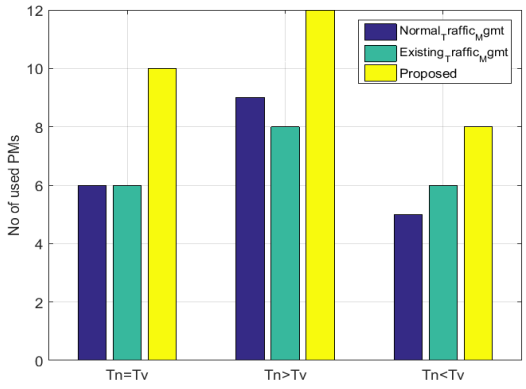

Fig. 2. Analysis of Number of used Physical Machines

The outcome in Fig.2 shows that normal traffic management uses less number of physical servers. Existing traffic management scheme is found to have similar usage of physical servers when normal traffic is equal to spiky traffic. However, in the case of difference, existing traffic management shows both lesser physical server usage (during $T_n>T_v$) and more physical server usage (during $T_n<T_v$). However, they exhibit more migration as compared to proposed system (Fig.3). This performance trend shows its attenuation pattern during the condition of $T_n>T_v$ and enhancement during the condition of $T_n<T_v$. The interesting finding is that by reducing the extent of live migration, the proposed system decreases its probability of downtime. Hence, the proposed system offers quite a less downtime and thereby exhibiting an efficient load balancing and server consolidation technique. The system, therefore, exhibits more enhanced performance by lowering down events of live migrations of VM. This outcome of lowered live migration will also have a positive impact on the capacity overflow parameter too.

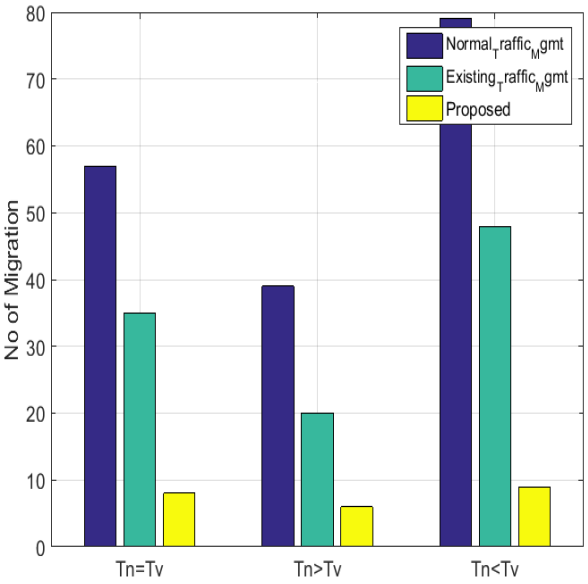

Fig. 3. Analysis of Number of Migrations

The primary intention of the proposed algorithm is to reduce the extent of the resources utilization that are kept conserved for the physical server while performing consolidation of the server and the cumulative system performance is ensured using probability theory. This will mean that a segment of time within which the collective traffic of the physical server is found to be more than its respective capacity is not higher as compared to the cut-off value of it. The proposed system applies usage of such cut-off values of capacity to resist overflow, and this phenomenon can significantly control an event of live migration thereby maintaining its capacity within very lower limits. In a nutshell, it will mean that if the capacity overflow can be controlled than live migrations of the VM can also be controlled too and hence capacity management over data center can directly influence the service quality. Fig. 4 showcases the analysis of the capacity for all the 3 different strategies.

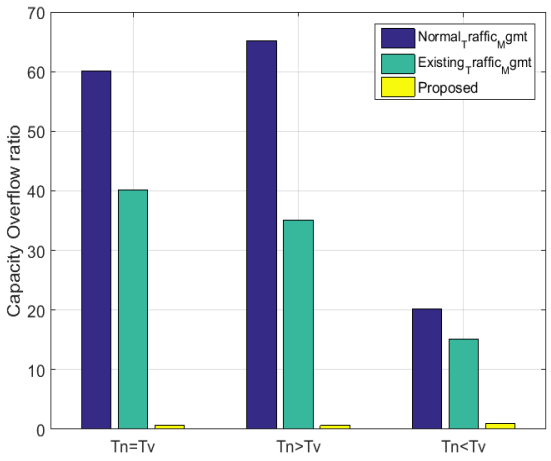

Fig. 4. Analysis of Capacity Overflow

A closer look into the graphical outcome of Fig.4 will show that proposed system to exhibit approximately 48% of enhanced performance as compared to normal traffic management and 30% improvement as compared to existing traffic management. To understand the link of capacity outcome with migrations, we consider an example where some of the physical servers may falsely declare itself as idle. It should be known that such false declaration of idle state is very common where an occupied physical server can be chosen as a target of migration. Such phenomenon will yield to the higher provisioning of physical serves resulting in iterative migration thereby causing downtime over the physical server in later stages. It also cost various resources associated with the VM to do the task scheduling under such forged cases of idleness. However, the extent of such cases is very low in the proposed system as it uses state-based transition along with probability theory that performs a minor computation of each and every resource and VMs along with its respective states. The complete testing was performed considering both stationary and changing values of traffic. To obtain convergence, the simulation study was carried out for multiple numbers of iterations individually in all the three cases of traffic pattern viz. i) $T_n=T_v$, ii) $T_n<T_v$, and iii) $T_n>T_v$.

The primary observation of the proposed system is the minimization of the number of physical servers as compared to normal traffic scene also shows more inclinations towards the adoption of such algorithms. The applicability, as well as need of such algorithm, is quite high for the massive transmission of the real-time data over the cloud. Effective allocation of the resource will further ensure a better balance between the user's request and service delivery. The outcome also showcases an effective VM migration management system along with a novel load balancing policy as well as server consolidation. Hence, a productive balance between the performance quality and utilization factor can be ensured by the proposed system. The proposed system can also be said to adopt the policy of multi-objective optimization policy where the objective function is to minimize the capacity overflow and live migration to retain a solid server consolidation scheme. The processing time of proposed technique for all the three different sates of traffic is found to be approximately 1.0576 seconds tested on core i3 machine with 64-bit Windows. When the operation environment changes than the accomplished outcome of the study only show 5% deviation as compared to stated numerical outcomes.

IV. Conclusion

With the increasing usage as well as the adoption of cloud computing, the technology consistently encounters critical challenges. One of the critical challenges that are discussed in this paper is resource management where the pivotal point of entire discussion was basically the role played by VM. In last 6 years, there has been enough number of research papers that has discussed various problems associated with VM including resource management, migration of VM, energy saving, security problems, etc. However, there is still a better scope of carrying out research work towards VM as still certain open research issues exist. The significant research issue is the lesser extent of computational modelling that focuses on the problem of live migration. There is a lot of difference between VM migration discussed in existing research work and live migration. To carry out live migration, the users will be required to be provided with the seamless delivery of services which is sustained by a higher allocation of various resources. The problems become worst if the time duration involved is more. Hence, live migration results in extensive resource usage and should be addressed properly. The proposed study presents a solution to this problem where an analytical modelling is introduced that maintains a good balance between resource management by lowering down live migration with increased capacity of VM. The outcome accomplished from the study was compared with normal traffic and existing system to find a proposed system outcomes existing system on increased use of physical servers, lower live migration, and increased capacity.

References

[1] X. Yang, "Principles, Methodologies, and Service-Oriented Approaches for Cloud Computing", *IGI Global*, 2013

[2] L. Tsai, W. Liao, "Virtualized Cloud Data Center Networks: Issues in Resource Management", *Springer*, 2016

[3] H. Saboowala, M. Abid, S. Modali, "Designing Networks and Services for the Cloud: Delivering business-grade cloud applications and services", *Cisco Press*, 2013

[4] Z. Mahmood, "Cloud Computing: Challenges, Limitations and R&D Solutions", *Springer*, 2016

[5] S. U. Khan, A. Y. Zomaya, "Handbook on Data Centers", *Springer*, 2015

[6] D. Mishchenko, "VMware ESXi: Planning, Implementation, and Security", *Cengage Learning*, 2010

[7] J. U. Gonzalez, S. P. T. Krishnan, "Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects", *Apress*, 2015

[8] C. McCain, "Mastering VMware Infrastructure 3", *John Wiley & Sons*, 2010

[9] N. L. S. da Fonseca, R. Boutaba, "Cloud Services, Networking, and Management", *John Wiley & Sons*, 2015

[10] D. Agrawal, S. Das, A.El Abbadi, "Data Management in the Cloud: Challenges and Opportunities", *Morgan & Claypool Publishers*, 2012

[11] S. Fiore, G. Aloisio, "Grid and Cloud Database Management", *Springer Science & Business Media*, 2011

[12] X. Zhu, Y. Zha, L. Liu, and P. Jiao, "General Framework for Task Scheduling and Resource Provisioning in Cloud Computing Systems", *40th IEEE Computer Society International Conference on Computers, Software & Applications*, 2016

[13] N. Kumara, S. Saxena, "Migration Performance of Cloud Applications-A Quantitative Analysis", *Elsevier-ScienceDirect- Procedia Computer Science*, Vol.45, pp.823 – 831, 2015

[14] A.T. Saraswathi, Y.RA. Kalaashri, S.Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", *Elsevier-ScienceDirect-Procedia Computer Science*, Vol. 47, pp.30–36, 2015

[15] T. Wood, K. K. Ramakrishnan, P. Shenoy," CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines", *IEEE/ACM Transactions On Networking*, 2014

[16] S. K. Panda, S. Nag and P. K. Jana, "A Smoothing Based Task Scheduling Algorithm for Heterogeneous Multi-Cloud Environment", *IEEE- International Conference on Parallel, Distributed and Grid Computing*, 2014

[17] J. Song, W. Liu, F. Yin, and C. Gao, "TSMC: A Novel Approach for Live Virtual Machine Migration", *Hindawi Publishing Corporation, Journal of Applied Mathematics*, 2014

[18] S.Selvarani, G.S. Sadhasivam, "Improved Cost-Based Algorithm For Task Scheduling In Cloud Computing", *IEEE International Conference on Computational Intelligence and Computing Research*, 2010

[19] A. Nahir, A. Orda, D. Raz, "Resource Allocation and Management in Cloud Computing", *IEEE International Symposium on Integrated Network Management*, 2015

[20] M-T Kao, Y-H Cheng, and S-J Kao, "An Automatic Decision-Making Mechanism for Virtual Machine Live Migration in Private Clouds", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[21] L. Wei, C. H. Foh, B. He, J. Cai, "Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds", *IEEE Transactions on Cloud Computing*, 2015

[22] W. Yue and Q. Chen, "Dynamic Placement of Virtual Machines with Both Deterministic and Stochastic Demands for Green Cloud Computing", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[23] S. Caton, C. Haas, K. Chard, K. Bubendorfer, O. Rana, "A Social Compute Cloud: Allocating and Sharing Infrastructure Resources via Social Networks", *IEEE Transactions On Services Computing*, 2014

[24] B. Rong C., H-F Tsai, and C-M Chen, "Empirical Analysis of Server Consolidation and Desktop Virtualization in Cloud Computing", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[25] C-T Yang, H-Y Cheng, and K-L Huang, "A Dynamic Resource Allocation Model for Virtual Machine Management on Cloud", *Springer Journal*, pp.581-590, 2011

[26] B. Perumal and A. Murugaiyan, "A Firefly Colony and Its Fuzzy Approach for Server Consolidation and Virtual Machine Placement in Cloud Datacenters", *Hindawi Publishing Corporation, Advances in Fuzzy Systems*, 2016

[27] V. Eramo, A. Tosti, and E.Miucci, "Server Resource Dimensioning and Routing of Service Function Chain in NFV Network Architectures", *Hindawi Publishing Corporation, Journal of Electrical and Computer Engineering*, 2016

[28] T. K. Sarker and M. Tang, "Performance-driven Live Migration of Multiple Virtual Machines in Datacenters", *IEEE International Conference on Granular Computing*, 2013

[29] K. Ye, Z. Wu, C. Wang, B. B. Zhou," Profiling-based Workload Consolidation and Migration in Virtualized Data Centres", *IEEE Transactions On Parallel And Distributed Systems*, 2013

[30] S. Andrilli, D. Hecker, "Elementary Linear Algebra", *Academic Press*, 2016