INTERNATIONAL JOURNAL OF ADVANCED STUDIES
IN COMPUTER SCIENCE AND ENGINEERING
IJASCSE VOLUME 6 ISSUE 9, 2017

09/30/2017

A review on: VM management in Cloud & Datacenter

Nikita Sharma
Computer Engineering
Govt. Women Engineering College, Ajmer, India

Sudarshan Maurya
Assistant Professor
Govt. Women Engineering College, Ajmer, India

***Abstract*—Cloud computing (CC) is very emerging and young technology for storing and accessing data and deploying application over the web/internet as a substitute of host thus user can access data from anywhere. Virtualization is the essential technology for enabling cloud computing and datacenters for cloud venders like Google, amazon, IBM, Microsoft, Citrix etc. Security, efficient utilization of resources, load balancing, power management, scalability, capacity planning, monitoring etc. are the most challenging issues that all organization face. They have to find the solution for those problems. This paper presents a survey for server consolidation and load balancing in data center and/or cloud. We focus on approaches used to consolidate Virtual machines (VMs) into the hosts. Load on the cloud/data center is dynamic which require VMs to be created and removed dynamically as per the load. In result which gives effective utilization of resources, energy efficient and approaches must fulfil the SLA and QoS requirement. VM management on cloud/data center is used to balance the load and replication of VM on physical machine to increase fault tolerance.**

***Keywords*—Server consolidation; Load Balancing; Virtualization; SLA; QoS; VM management.**

I. INTRODUCTION

Cloud computing is web-based computing technology. It delivers the services over the internet or Intranet and a service of application, computing and storage is provided on pay per use. Below are mostly quoted definitions of the cloud computing:-

From a paper (by Buyya et al., 2009) "A cloud is a type of parallel and distributed system that consist of a pool of interconnected and Virtualized computers, which are dynamically provisioned and presented as one or more unified computing resources(s) based on services-level agreements (SLA) established over negotiation between the service provider and consumer"[1].

From a paper by NIST (Mell and Grance,2011) "Cloud Computing configuration is a model for accessing a shared pool of computing resources(such as networks, servers, storage, applications and services) to the ubiquitous, convenient, on-demand network which can be rapidly provisioned and minimal management effort or services provider can be released with the conversation. This cloud model is composed of five essential features, three service models, and four deployment models [2]".
Some of the Cloud management products are:-VMware Capacity Planner, CapacityIQ, IBM Websphere cloudburst, Novell PlateSpin Recon, and Lanamark suite [16] etc. Virtualization is the essential technology of CC and Datacenters. It's a technique to run many operating systems simultaneously on a host machine. It allows resource multiplexing, live migration, server consolidation, energy management, VM resizing, VM scaling cluster maintenance and load balancing. Typically Datacenters provides services of homogeneous nature whether cloud provides services of heterogeneous nature, but core virtualization technology is same in both.

The Term Infrastructure as a "cloud" from which the application as a service are available from everywhere on request of business and users [1]. Infrastructure as services (IaaS)-cloud providers provide the resources on-demand from the vast pool of resources equipped on Datacenters [7]. User requirement varies with the time therefore Cloud datacentres are must be more flexible, secure and efficient to run the complex workload demand and different applications. We can say datacentre are the resource providers. Cloud Infrastructure refers to the hardware and software component such as - Virtual machines, server storage, load balancer, networks etc. Example of Infrastructure as a services (IaaS) service suppliers are:-Amazon EC2 [3], Microsoft Azure [4], Google Compute Engine [5] etc. Host system is a system on a network, which provides the services to user or other computer on that network. In our paper we are focusing on the Load Balancing of VMs on the host within the cloud. In load balancing VM can migrate from one host to a different less busy host. If one host fails then we can recover it by reinstalling the same VM on another host. VM migration is used for balance of load on host in the cloud.

There are two types of VM migration (VMM) [10]:-

- **Online(warm/Live) migration:-** Basic plan of online migration preliminary proposed by "Christopher Clark" [8].In online migration, running VM can be transferred across to the one host to additional different host without disturbing the services; user can't predict the interruption of services [9]. In this process, Firstly VM is put off, then it is relocated and lastly VM is restarted at targeted host. We can dynamically balance the workload [14].
- **Offline** (**Cold/non-live) migration: -** In cold migration, running VM can be transfer to the one host to another host; during this transfer user can predict the interruption in services.

There are many reasons to do the VM migration [10]:-

1) Flexibility to handle Dynamic Load: Workloads are dynamically changing with the time in server. There is frequent spike of request. Migration can adjust both additional workload and reduced workloads. It can reduce the problem of over-provisioning and under-provisioning.
2) Maintenance: Schedule maintenance, it gives some downtime for the user of the server. For some time, load is transfer to the VM and when maintenance is complete then load is brought back to the host VM.
3) System Failure: If some unexpected fault occurs in server which gives unscheduled server downtime in result, problem can be solved by the VM migration, so that user can experience high availability of application at all time.
4) Disaster Recovery: Recovery from host failure, it is possible for migrating operating system and application from older server to the newer server easily and without disrupting the services.
5) Mobility: if customer having internet connections so Customer can access data and share data anytime from anywhere.
6) To optimize physical resource utilization: We can appropriately move idle VM, near-idle VM, or VM with currently less-critical guest workloads together on a smaller or less powerful machine.

II. SYSTEM MODEL

Typical Datacenter Architecture along hosts is shown in figure (1).

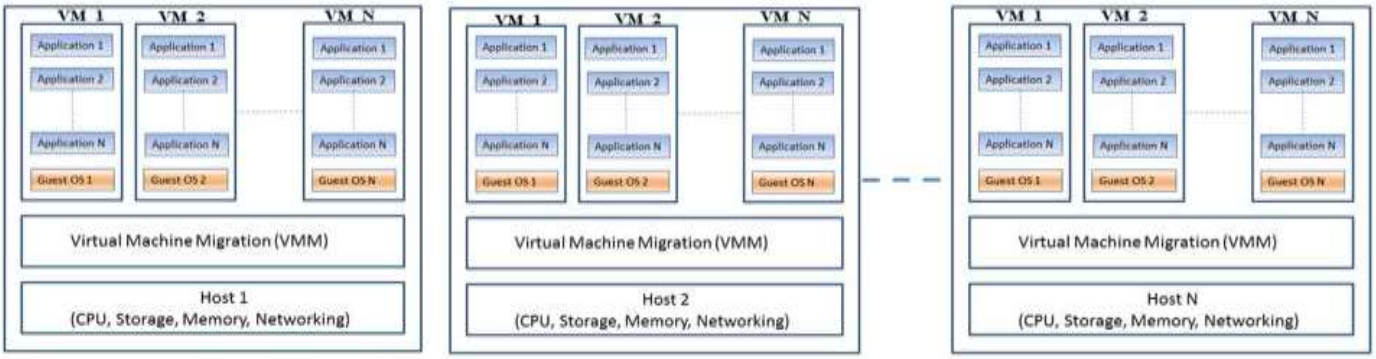

Figure 1 Hosts on Datacenter or cloud

Datacenter contain multiple numbers of hosts which is used to fulfil the customer need. Virtual machine monitor (VMM) can virtualize the whole physical machine. VMM runs one or more virtual machine known as host machine and each VM is known as Guest machine. It is software which configures and manages virtualization host, networking, and storage resource in manner to create, edit, start or stop VM. The Application and Guest Operating System (OS) are available in a Virtual Machine (VM).To meet the customer's demand; Host can face overloading and underloading conditions so that the server can be used efficiently. There are various methodologies to balance the upcoming load, some of them which are given below:-

A. Server consolidation

Virtualization is the abstraction of the physical resources into virtual resources that decoupled from underlying hardware's. Server Consolidation is a method to use the computer servers in resourceful manner to reduce the total number of required servers. This method developed in response to the problem of "server sprawl", it is a condition in datacentres which gives under-utilized servers which consume more storage and resources, and waste energy in result. A business organization peruses financial saving by using both server and storage consolidation. **Steinder, Malgorzata et al., 2007**, propose a method for autonomic management of heterogeneous workload on any server [10]. The problem is based on the scheduling of web and non-interactive batch workloads across a cluster of VMs and servers to meet the respective SLAs goals. They combine the "flow control and dynamic

placement techniques" with the "job scheduling" to manage the heterogeneous workload effectively. In this approach, migrations are often used for transfer the image to a different node, once required. This work not ideal for the cost optimization and for performing virtualization actions and doesn't save power by switching off the unutilized machines. This system is capable of using many types of virtualization for many requests. **Van et al., 2009,** discourses the problem of autonomic resource management to manage the dynamic placement of virtual machine supported the Service Level Agreement (SLA) and energy consumption [14]. Resource manager used to optimize the utility function and utility function used to determine the QoS (quality of the service) by using the completion of the demand, resource management costs. Self-optimization can be achieved by a mixture of utility function and constraint satisfactory problems (CSP) (VM provisioning and VM packing problem are expressed as two CSP). In this approach live migration is used. And it is said that VM's migration cost is similar to the quantity of memory allotted to VM. They did not consider the cost of migration.

B. Load balancing

Load Balancing ensure the evenly distribution of workload and applications demands by allocating resources among multiple computer, networks or servers to complete the running task on the time [33]. Load can be CPU capacity, memory, network or delay load. VMs will be transferred from the over load VM to less loaded VM to avoid the overburdening, to improve resource utilization and to enhancing the overall system performance. It helps to unbiased allocation of resources to achieve high resource utilization and proper load balancing.in daily life ,user can face the problem of delay, system time out ,and long to response etc. because many web sites doesn't use the load balancing.

Types of load balancing:-

a) ***Static Load Balancing Algorithm****:* This type of algorithm is non-preemptible [33]; they are efficiently works for the static and homogenous workloads. It requires the prior knowledge of system resources like node processing power; capacity, memory etc. reduce the execution time and communication delay between nodes. Example of Heuristic based static algorithms: -'First-come-First-Serve (FCFS)', 'Round-Robin (RR)'[31], 'Randomized Algorithms',' Central Manager Algorithm', etc. **Sotomayor et al., 2008,** discuss a leasing based architecture, to handle the scheduling of combination of both preemptible best-efforts for predicting various run-times overhead involved in using virtual machines and support the efficient advance reservation leases [12].

b) **Dynamic Load Balancing Algorithm**: This type of algorithm is preemptible [33], it can works for dynamic and heterogeneous workloads. It can consider various parameters of system resources prior or during the running time. It requires the communication between the nodes. Examples: - Ant colony algorithm, local Queue, Central Queue etc.**Hu et al.,2009,** developed an efficient and effective algorithm for allocation of the resources which is applicable to autonomic resource management, in result algorithm gives the minimum number of required server and the probability dependent policy (PDP) which is designed to maximize the probability of meeting a given response time goal and more than FCFS and head-of-the-line (HOL) priority in terms of requiring the minimum number of servers [15]. HOL, perform the task which having the higher priority. They show PDP require less host to meet the SLA. This Technique can be used to develop heuristic method when more than two classes are allocated to application. The performance of shared allocation (SA) with the FCFS scheduling also investigated. **Nathuji et al., 2010,** developed a "QoS-aware control framework" for multi-core Cloud Servers which are a feedback-based scheme that manages effective and dynamic resource allocation and QoS for dynamic workload [19]. They use "multi-input multi-output (MIMO) model" to capture the performance interference effects to run a closed loop resource management controller. In addition, Q-cloud increases the resource utilization by 35%. In this paper they don't investigate the performance metrics for dynamic workload consolidation using live migration techniques.

III. RELATED WORK

A. Classification based on VM Consolidation Type

1) ***Static Consolidation****:* In static consolidation, size and placement of VMs on PMs is determined statically when jobs arrives and not change with the period of time [10].The advantage of this system is in the process of batch job and application with stable and forecasted demand.And drawbacks of this approach is resources are allocated according to the peak of demand so most of the resources are waste and this problem generate over-provisioning means service provider; provide the resources beyond the maximum need of user.

2) ***Dynamic Consolidation:*** In dynamic consolidation, VM capacity is dynamically changes according to the workload fluctuation within the specific time interval. By using Dynamic Consolidation we can improve the effectiveness and reduce the performance and capacity overhead. **Wang et al., 2007,** argues that demand is frequently changes with the time thus service

provider making an attempt to fulfil the SLA there is need to reallocate the resources or reschedule the resources, in result that gives the performance overhead, capacity overhead and actuation delay, to solve this overhead they provide the concept of controlling the system which can dynamically allocate resources with the fluctuating user demand and requirement. They show the feasible overhead of the dynamic allocation technique in compare to static allocation technique. In result, Due to the dynamic allocation of CPU capacity, they saw a decrement in work-load performance and lack of system capacity. Compared to the OpenVZ, Xen system has a high performance and capacity overhead for both computational and transactional workload. Virtualization overhead increases with the no of virtual containers. More rapid controller response and fewer capacity and performance overhead will be seen in result [11].

B. Classification based on targeted systems

Soundararajan et al., 2010, is examine the information from real-world virtualized organization to distinguish common management workflows and evaluate the impact on resource use within the datacenter, they analyze how to datacenter scale as the requirement change and discuss the several trade-offs in the design of a datacenter of virtualized environment [17]. It is a storage based virtualization, like Storage area network (SAN) provides different logical view of physical storages. Usually it is the single consolidation on the multiple physical devices. They show management workload scales as the server's calculation power will increase and will be much faster with increasing production of multi-core processors.

Table 1 Comparative Study

References	Year	Time of decision making strategy			Workload		Goals						Migration	Virtualization Platform	Performance Evaluation of				Evaluation Strategy		
		Static	Dynamic	Dynamic with load prediction	Web based application	non- interactive batch jobs	Energy Efficient	SLA And QoS	Capacity and Performance overhead	Maximization Profit	Resource utilization	Cost Effective			CPU	Bandwidth	Memory	Network	Simulation	Experiment	Emulation
Steinder et al. [6]	2007		✓		✓	✓		✓			✓		✓	Xen	✓		✓		✓		
Wang et al. [7]	2007		✓		✓	✓		✓	✓					Xen OpenVZ	✓					✓	
Sotomayor et al.[8]	2008			✓		✓			✓		✓		✓		✓	✓	✓	✓	✓		
Quiroz et al. [9]	2009		✓		✓	✓			✓		✓	✓	✓		✓		✓		✓		
Van et al. [10]	2009			✓	✓	✓		✓				✓	✓		✓		✓		✓		
Meng et al. [11]	2010			✓	✓	✓		✓			✓		✓	VMware	✓		✓		✓		
Singh et al. [14]	2010			✓	✓	✓		✓		✓	✓		✓	Xen	✓	✓	✓	✓		✓	
Zhang et al. [16]	2010		✓		✓	✓			✓		✓		✓	VMware	✓			✓		✓	
Garg et al. [17]	2011		✓		✓	✓		✓	✓		✓		✓		✓				✓		
Kim et al.[18]	2013		✓		✓		✓	✓				✓	✓	Xen	✓		✓		✓		
Casalicchio et al.[19]	2013			✓	✓	✓		✓	✓	✓	✓		✓		✓		✓		✓		
Garg et al.[20]	2014			✓	✓	✓		✓			✓		✓		✓				✓		
Shahzad et al.[21]	2015	✓			✓	✓	✓	✓	✓			✓	✓		✓			✓	✓		
Wei et al.[22]	2015			✓	✓	✓					✓		✓		✓		✓		✓		
Sampaio et al.[23]	2015	✓			✓	✓	✓	✓	✓				✓						✓		
Antonescu et al.[24]	2016			✓	✓			✓	✓		✓		✓								

1) ***Arbitrary:*** **Singh et al., 2010** argues that the non-stationarity in web application workloads, due to which there are changes in mix of request over time, will have a major effect on overall processing demands placed on datacenters servers [18]. To ensure a minimum level of performance and meet the SLA, the system will need to dynamically match the assigned capacity for such fluctuations in the workload. Reducing over-provisioning in this technique gives efficient resource utilization and reduces the SLA violation due to under-provisioning. **Zhang et al., 2010**, focused on managing and quickly reassignment of the resources to supports the varying demand for these applications [20]. They proposed the concept of ghost VM approach for rapid resource allocation, ghost VM detached from the internet and additional capacity not deployed until the capacity is not needed. They focus on the capacity and utilization measures on CPU consumption. It works well with the workload fluctuation and gives efficient utilization of resources. request mix which is fluctuate with the time that can be known as bursty workload [18] ,to overcome from this bursty workload we can consider the ghost VM [20] as a solution.
2) ***Homogeneous Workload:*** Which have fixed number of parameters as following-amount of memory, Number of CPUs required, amount of local storage and buffer size of input and output files. **Steinder and Malgorzata et al., 2007,** shows the utility of the server virtualization technology within the management of homogenous workload, in the case of non-interactive batch jobs (workload)[10]. **Hu et al., 2009,** Shows the comparative evaluation of shared allocation and dedicated allocation under FCFS scheduling [15]. And also gives a heuristic algorithm which defines the resource scheduling strategy. **Casalicchio et al.,2013,** they adopt to a heuristic solution based on hill climbing search techniques with fixed parameter i.e. MaxRestarts searches, to minimize the possibility of hill climbing search to be stuck in a local optimum [23].
3) ***Heterogeneous Workload:*** Heterogeneous computing represents systems that use more than same type of processor or core, not only by adding the same type of system processors, but also improving performance and energy efficiency, but with specific processing capabilities to handle specific tasks by adding different coprocessor's [6].This type of system supports heterogeneity, where user must be able to make a request at instance of time with the configuration of system to fulfil the requirement. User can select services through the user interface using predefined request types or select additional services as per the need. User requests send to the scheduler, selecting computing resources according to user requests [30]. **Steinder and Malgorzata et al., 2007,** presents a system that enables any server machine to collocate heterogeneous workload, thus decreasing the unprocessed resource allocation [10]. **Garg et al.,2011,** proposed An efficient resource technique will require automatic allocation of heterogeneous resources and the minimum resource requirement can be fulfilled according to their SLA, and for additional resource requirement, additional virtual machine needed[21]. With migration policy, they try to optimize the use of minority VM by migrating and consolidating, which result in a large number of migrations. Migrating overhead batch causes unnecessary delay in performance, which is about 45% of successful completion before deadline. This approach is quite good and easily applies in a real-time cloud computing environment. They use single core CPU architecture and have memory conflicts. They ensure the SLA requirement meets and maximize the utilization of resources and revenue. **Garg and Saurabh Kumar et al., 2014,** solve the problem of efficient resource allocation within the datacenter running heterogeneous resources demands including both transactional and non-interactive batch jobs [24]. They proposed "admission control and scheduling mechanism" which enhance the resource utilization, revenue and also ensure the SLA and QoS have to be completed. They implement a mixture of workloads for different types of SLA penalties and enhanced resource provisioning and deployment of datacenters. They don't consider the resource starvation problem. **Wei, Lei, et al., 2015,** propose a heterogeneous resource allocation approach, called "Skewness-avoidance multi-resource allocation (SAMR)", to ensure the heterogeneous workload is allocated properly on the host in order to avoid skewed resource utilization [26]. Algorithm improves the resource uses and fixes the resource starvation problem. **Sampaio et al, 2015,** addresses the problem of resource allocation among the datacenter that runs heterogeneous resource, mainly demands CPU and network intensive application [27]. Propose energy and performance efficient enforcing mechanism to meet the QoS requirement, including performance deviation estimator and a scheduling algorithm. Proposed mechanism reduces the performance interference (i.e. sharing and concatenation of other resources) and power consumption but doesn't support dynamic VM re-sizing.

C. Classification based on the workloads

1) ***Transactional Application****:* provide a set of VM with the fluctuating capacity and hourly charges, allowing the user to select according to their needs[24]. Resources are equal to users peak demand. It includes Real-time applications/web-applications, whose demands for resource can vary over time depending on incoming request of applications.

2) ***Non-interactive batch jobs:*** is the execution of the series of the jobs without any manual involvement[24]. These types of jobs are sorted by threshold time,for which a batch job can be delayed without any penalty. It includes HPC-applications.

3) ***Mixed Workload:*** combination of both transactional application and non-intractive batch jobs known as workload mix[10] [11][12] [18].

D. ***Classification based on Fault tolerance***

1) ***Overheads Rejection:*** **Wang et al.,2007**, observed CPU performance overhead of workload performance and capacity overheaad due to dynamic allocation [11].

2) ***Capacity and Performance Management:***

a) ***Resources Utilization****:* **Quiroz et al., 2009**, explores autonomic approaches for optimizing VM provisioning for heterogeneous workload on enterprises grids and clouds [13]. Goal of paper [13] is, to improve the utilization, reduce the overprovisioning cost by using decentralized online clustering (DOC) and also deals with under provision problem. VM must be managed in run-time, according to data center capacity, load-balancing and long-term service constraints. To manage the VM ensure that the SLA and QoS must meet. **Meng et al., 2010,** promotes a "joint VM-provisioning approach"; in this approach various VM's are provisioned and consolidate on the basis of their aggregated estimated capacity requirement [16]. A joint VM sizing algorithm which calculates the total required capacity for multiplexed VM's. VM selection algorithm for joint sizing that detects compatible VM combinations (known as super VM) according to demand pattern for being consolidated and super-VM provisioned jointly to save high capacity. Approach gives the idea how can optimally place the VMs to leads to the higher utilization. There are two improvements in VM selection technique:-1) the datacenter size increase and correlation matrix scalability and 2) Merely recognize sets of compatible VM's.

b) ***Scalability:*** **Antonescu et al., 2016**, proposed two novel VM-scaling algorithm dedicated on DEIS (distributed enterprises information system) system can be used by cloud infrastructure, so that the most appropriate scaling condition can be detected, which consistently showcase the performance model of distributed application derived from constant- workload benchmarks and allow cloud management system (CMS) to ensure the performance and QoS requirement defined in SLA [28]. The Algorithm maintains 100% SLA compliance rate and improves scalability and improves the management efficiency.

c) ***Power Consumption****:* **Kim et al., 2013,** propose a dynamic power management solution for server hosting these new scale-out applications [22]. Solution jointly utilize homogenous server consolidation and voltage and frequency (v/f) scaling, considering the characteristics of scale-out application, especially correlation among VMs. "Dynamic V/f scaling" is used to reduce the power consumption to meet the QoS requirement. All co-located VMs share cores that give negligible power-degradation. To find the optimal sets of VMs they propose "Correlation-aware VM allocation technique" which is based on first-fit decreasing heuristic technique. According to author proposed approach provide up to 13.7 % power-saving and up to 15.6% improvement of QoS level in compare to VM placement solutions.one problem is faced by proposed approach, if more servers needed to achieve the QoS level(here QoS is scalability and reliability), which leads to higher power consumption.

d) ***Revenue:*** **Casalicchio et al., 2013**, present an autonomic method to design of Self-Optimizing cloud provider. Paper aims to improve the profits and capacity overhead, meet the SLA and VM migration constraints [23]. They present 'a heuristic solution called near optimal (NOPT)', to this NP-hard problem and compare the result with the Best-fit allocation strategy. In addition, they provide a formulation that decreases the VMs Migration but not an optimal allocation, VMs is migrating during the admission control phase, by using hill climbing search techniques. The system provides 45% improvements in the average revenue and more opportunity to rearrange VMs to achieve better revenue.

e) **Energy Efficient:** Energy-saving can be achieved by the continuous VM consolidation according to current resource requirements [32].**Sampaio et al., 2015**, Fulfil signed SLA's of real-time computing environment and also reducing energy costs by 21% [27].

IV. CONCLUSION

Cloud computing deals with the Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) without being affected by the location. In this paper we are trying to construct the understanding of the concept of server consolidation and load balancing and here we are showing the various classification of the server consolidation based on the various aspects and concept of load balancing in Datacenter at the VM level. In comparative study we analyze which algorithm is best for the load balancing. Jobs are requested from the user and service provider provides the services according to the user need. Here each job is assigned to Virtual machine and Service Provider always trying to increase revenue and give the best services to users according to their signed agreement.

V. FUTURE SCOPE

According to our survey we found all service providers face the problem of Energy efficient, resource utilization, reducing the migration cost, power consumption, quality of services etc., is still a biggest problem for every cloud provider. Those approaches can be extended and further simulate for efficient result.

References

[1] Buyya, Rajkumar, et al. "Cloud computing and emerging IT platforms, Vision, hype, and reality for delivering computing as the 5th utility." Future Generation computer systems 25.6 (2009): 599-616.APA
[2] Mell, Peter, and Tim Grance. "The NIST definition of cloud Computing."(2011).
[3] Amazon Elastic Computing Cloud, aws.amazon.com/ec2
[4] Azure, https://azure.microsoft.com/en-in/
[5] Google Cloud, https://cloud.google.com/compute/
[6] Shelepov, Daniel, et al. "HASS: a scheduler for heterogeneous multicore systems." *ACM SIGOPS Operating Systems Review* 43.2 (2009): 66-7.
[7] Armbrust, Michael, et al. "A view of cloud computing." Communications of the ACM 53.4 (2010): 50-58.
[8] Clark, Christopher, et al. "Live migration of virtual machines." Proceedings of the 2nd Conference on Symposium on Networked Systems Design and Implementation-Volume 2. USENIX Association, 2005.
[9] Kapil, Divya, Emmanuel S. Pilli, and Ramesh C. Joshi. "Live virtual machine migration techniques: Survey and research challenges." Advance Computing Conference (IACC), 2013 IEEE 3rd International. IEEE, 2013.
[10] Steinder, Malgorzata, et al. "Server virtualization in autonomic management of heterogeneous workloads." Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on. IEEE, 2007.
[11] Wang Z, Zhu X, Padala P, Singhal S. Capacity and performance overhead in dynamic resource allocation to virtual containers. In: Proceedings of the 10th IFIP/IEEE international symposium on integrated network management. Munich, Germany; 2007.
[12] Sotomayor B, Keahey K, Foster I. Combining batch execution and leasing using virtual machines. In: Proceedings of the 17th international symposium on HPDC. Boston, MA, USA; 2008
[13] Quiroz A, Kim H, Parashar M, Gnanasambandam N, Sharma N. Towards autonomic workload provisioning for enterprise grids and clouds. In: Proceedings of 10th IEEE/ACM international conference on grid computing. Melbourne, Australia; 2009.
[14] Van HN, Tran FD, Menaud J-M. SLA-aware virtual resource management for cloud infrastructures. In: CIT '09: proceedings of the 2009 ninth IEEE international conference on computer and information technology. Xiamen, China; 2009.
[15] Hu Y, Wong J, Iszlai G, Litoiu M. Resource provisioning for cloud computing. In: CASCON '09: Proceedings of the 2009 conference of the Center for Advanced Studies on Collaborative Research, Ontario, Canada; 2009.
[16] Meng X, Isci C ,Kephart J ,Zhang L ,Bouillet E ,Pendarakis D. Efficient resource provisioning in compute clouds via VM multiplexing. In: Proceedings of the 7th international conference on autonomic computing, Washington, USA;2010.
[17] Soundararajan V, Anderson J. The impact of MNGT. Operations on the virtualized datacentre. In: Proceedings of the 37th annual international symposium on computer architecture. France; 2010.
[18] Singh R, Sharma U, Cecchet E, Shenoy P. Autonomic mix-aware provisioning for non-stationary data center workloads. In: Proceedings of the 7th international conference on autonomic computing. Washington, USA; 2010.
[19] Nathuji R, Kansal A, Ghaffarkhah A. Q-clouds: managing performance interference effects for qos-aware clouds. In: Proceedings of the 5th European conference on Computer systems (EuroSys 2010). Paris, France; 2010.
[20] Zhang W, Qian H, Wills C, Rabinovich M. Agile resource management in a virtualized data center. In: Proceedings of Ist joint WOSP/SIPEW international conference on performance engineering. California, USA; 2010.
[21] Garg S, Gopalaiyengar S, Buyya R. SLA-based resource provisioning for hetero- geneous workloads in a virtualized cloud datacenter. In: Proceedings of the 11th international conference on algorithms and architectures for parallel Processing, Melbourne, Australia 2011.
[22] Kim J, RuggieroM,AtienzaD,LederbergerM.Correlation-awarevirtualmachine allocation forenergy-efficient datacenters.In:Proceedingsoftheconferenceon design, automationandtestinEurope.Ghent,Belgium;2013.
[23] Casalicchio E, Menascé DA, Aldhalaan A. Autonomic resource provisioning in cloud systems with availability goals. In: Proceedings of the 2013 ACM cloud and autonomic computing conference, Miami, FL, USA; 2013.
[24] Garg, Saurabh Kumar, et al. "SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter." Journal of Network and Computer Applications 45 (2014): 108-120.
[25] Shahzad, Khyyam, Arif Iqbal Umar, and Babar Nazir. "Minimizing SLA Violations and VMs Migration to Reduce Network Load in Cloud Data Centres." Ist Multi-Disciplinary Research Conference 12th September 2015, University of Sargodha Mandi Bahauddin Campus Punjab Pakistan.
[26] Wei, Lei, et al. "Towards efficient resource allocation for heterogeneous workloads in IaaS clouds." IEEE Transactions on Cloud Computing (2015).
[27] Sampaio, Altino M., Jorge G. Barbosa, and Radu Prodan. "PIASA: A power and interference aware resource management strategy for heterogeneous workloads in cloud data centers." Simulation Modelling Practice and Theory 57 (2015): 142-160.
[28] Antonescu, Alexandru-Florian, and Torsten Braun. "Simulation of SLA-based VM-scaling algorithms for cloud-distributed applications." Future Generation Computer Systems 54 (2016): 260-273.
[29] Kaur, Rajwinder, and Pawan Luthra. "Load balancing in cloud computing." Second Symposium on Cloud computing. 2012.
[30] Crago, Steve, et al. "Heterogeneous cloud computing." Cluster Computing (CLUSTER), 2011 IEEE International Conference on. IEEE, 2011.
[31] Pasha, Nusrat, Amit Agarwal, and Ravi Rastogi. "Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment." International Journal of Advanced Research in Computer Science and Software Engineering 4.5 (2014).
[32] Beloglazov, Anton, and Rajkumar Buyya. "Energy efficient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing. IEEE Computer Society, 2010.
[33] Chaczko, Zenon, et al. "Availability and load balancing in cloud computing." *International Conference on Computer and Software Modeling, Singapore*. Vol. 14. 2011.