I.J. Information Engineering and Electronic Business, 2017, 1, 28-35
Published Online January 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2017.01.04

Optimization Techniques for Resource Provisioning and Load Balancing in Cloud Environment: A Review

Amanpreet Kaur
Research Scholar, IKGPTU, Jalandhar
Email: er.amanpreet14@gmail.com

Dr. Bikrampal Kaur
Professor, Chandigarh Engineering College, Landran (Mohali)
Email: mca.bikrampal@gmail.com

Dr. Dheerendra Singh
Associate Professor, CCET, Chandigarh
Email: professordsingh@gmail.com

Abstract—Cloud computing is an emerging technology which provides unlimited access of versatile resources to users. The multifaceted and dynamic aspects of cloud computing require efficient and optimized techniques for resource provisioning and load balancing. Cloud monitoring is required identifying overutilized and underutilized of physical machines which hosting Virtual Machines (VMs). Load balancing is necessary for efficient and effective utilization of resources. Most of the authors have taken the objective to reduce the makespan for executing requests on multiple VMs. In this paper, a thorough review on scheduling and load balancing techniques has been done and different techniques have been analyzed on the basis of SLA Violations, CPU utilization, energy consumption and cost parameters.

Index Terms—Virtual machines, cloud monitoring, load balancing, resource provisioning, scheduling and optimization techniques.

I. Introduction

Cloud computing is a combination of distributed, parallel, multi-tenant computing model based on different technologies like virtualization, grid, utility and autonomic computing. Today is the era of IT and internet which have converged various services related to hardware (servers, networks, storage etc), software (security and finance monitoring services, testing modules, CRM modules, ERP software etc), platforms and communications converge to a single window. All these services collectively called CLOUD. Cloud Computing represents the outsourcing of these services to an external service provider and involves three basic services [2]:

1. Resource Discovery- Among the pool of cloud resources, the one which satisfies the requests is selected.
2. Monitoring- it is important to monitor the usage of cloud resources so that they could be utilized efficiently and effectively.
3. Load Balancing- Each host must have balanced workload to ensure that hosts should neither be over-utilized nor under-utilized.

In literature, huge number of approaches is proposed for above mentioned services. These approaches vary on the basis of their static / dynamic nature, centralized / distributed algorithms, hierarchical and workflow dependent techniques. Various cloud monitoring tools are available as commercial or open source like Amazon's Cloud Watch, MS-windows Azure's Fabric Controller, Eucalyptus, Open Nebula,

Nimbus, Cloud stacks ZenPack etc. which monitors, maintain and provision the cloud resources to the application requests. They monitor the performance of cloud services and allocate resources depending upon the resource capability to handle the requests.

Virtualization is the core technology used in cloud computing which allocates multiple virtual machines (VMs) to physical hosts. Hypervisors or virtual machine monitors are responsible for assigning the various VMs onto a single physical machine, hence improves the physical machine utilization. It also decreased the number of hosts required to execute same number of applications as are needed in non-virtualized environment [5]. Hosts at a single location form a datacenter and datacenter at distributed locations collectively serve the requests from multiple cloud customers.

In section II different aspects of cloud provisioning like scheduling, load balancing, cloud monitoring has been discussed along with the important parameters on the basis of which load is distributed among nodes. Related work is presented section III with extensive

literature review and the earlier work done has been analyzed on the basis of time, cost, SLA violations and energy efficiency parameters.

II. Cloud Provisioning

The pool of resources is shared among multiple consumers while ensuring maximum utilization of resources (on service provider side) and minimum operating cost (on customer side). To harvest from the benefits of cloud computing, it is important to provision the cloud resources and map tasks to cloud resources efficiently and optimally. The process of deploying the application services on the resources provided by the cloud is called cloud provisioning. It involves two steps:

1. VM Provisioning- hosting VMs on the physical servers in the datacenters of Public/Private clouds.
2. Application Server Provisioning- mapping and scheduling applications/ requests onto the VMs by following the load balancing approaches.

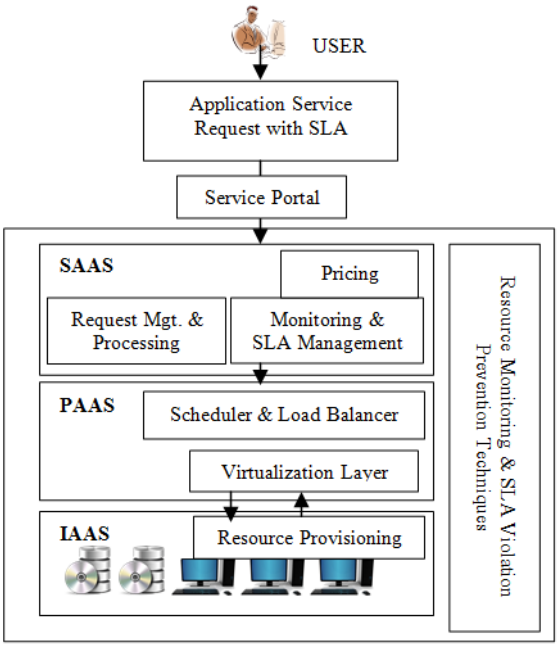


Fig.1. Cloud Resource Provisioning and Deployment Model [9]

A. Scheduling

In cloud taxonomy, Scheduling refers to mapping the tasks to suitable resources to optimize their utilization. Tasks must be allocated efficiently to VMs with minimum effort and communication delay. The resources are allocated and utilized by VMs hosting on physical machines. In cloud computing, the demands of user are highly dynamic in nature and multi-tenancy requires isolating different users from each other and from the cloud infrastructure. The providers have to follow the SLAs and ensure the Quality of Service (QoS) requirements as mentioned by the customers.

Many researchers have proposed and found sub-optimal solutions towards this problem. The key parameters that are tried to optimize are minimizing makespan (total time to complete the last task on a particular resource), response time, execution time, energy consumption, operating cost and maximizing resource utilization (CPU, memory, disk, network bandwidth).

In literature, task scheduling in cloud computing has been done using different optimization techniques to obtain best or sub-optimal schedules. In cloud environment, the metaheuristics optimization algorithms mainly used for the job scheduling are Ant Colony optimization algorithm(ACO), Particle Swarm Optimization(PSO), Genetic Algorithm(GA), Cuckoo Search (CS), Ordinal Optimization(OO), Bee Colony optimization(BCO), Multi-objective Genetic Algorithm (MO-GA) and using combination of optimization algorithms to obtain hybrid optimization techniques.

B. Load Balancing

When tasks are scheduled on VMs hosting on physical nodes, there might arise the situation that some of the nodes are over-utilized while others are under-utilized. So Load balancing is important for optimized use of cloud resources (processors, memory, disks) and to provide high performance of the machines. The objective of Load Balancing is to distribute jobs or tasks among available resources so that the relative imbalance is minimized, and the utilization of resources is almost equal. Workload Balancing as defined by Raj Kumar et al. (2004, 2006, and 2007) is the minimization of total relative imbalance which is defined as the ratio of the difference between maximum completion time on all machines and individual completion time on a given machine, and the maximum completion[4].

To distribute the dynamic cloud workload (tasks) evenly among the processing nodes, load balancing is done. Load balancing is required to avoid the occurrence of overloaded and under loaded nodes. However, the load can refer to CPU utilization, memory or storage usage or it can network load. Most the researchers have taken CPU utilization as the primary factor to take load balancing decisions. Load Balancing techniques are categorized as in fig. 2:

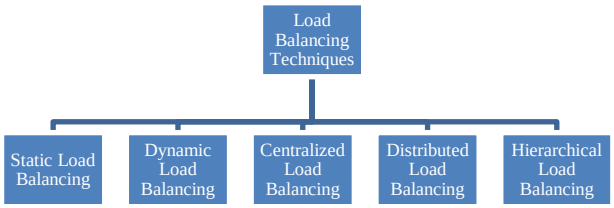


Fig.2. Different Techniques of Load Balancing

Static load balancing algorithm works in stable and homogeneous environment which is based on the previous knowledge about the node attributes (processing power, memory, storage etc). No changes are accepted in node statistics during run time. Dynamic Load Balancing

(DLB) considers the varying operating conditions of a heterogeneous environment. It works on the run-time statistics of the node which require continuous monitoring of load changes during execution. They are accurate, complex and have high time complexity and are used to transfer load from heavily loaded host to lightly loaded host with limit on the number of migrations. Central Load Balancing Decision Model(CLBDM), Round Robin Load Balancing algorithm, Enhanced Map Reduce algorithms are based on static load balancing technique whereas Index Name Server (INS), Exponential Smooth Forecast based weighted least connection (ESWLC), dual direction downloading Algorithm(DDFTP) and Load Balancing Min-Min (LBMM) use dynamic technique to balance load [10]. Further, static and dynamic load balancing algorithms follow centralized or decentralized approaches. In centralized approach, a single node or server maintains the details of the nodes and the network. It updates this information periodically and centrally responsible for all load balancing activities. This approach is suitable for small networks, but may suffer from bottleneck and single point failure.

Distributed load balancing requires each processor in the network to keep a copy of the global state of the nodes and network in their own database. All nodes are responsible for load balancing. However, distributed load balancing can be sender initiated, receiver initiated or symmetrically initiated by both sender and receiver. In hierarchical approach, multiple nodes are arranged in tree structure such that the parent node broadcast the node/network performance information to their child nodes.

C. Cloud Monitoring

To keep check on the usage of the cloud resources, performance, Auditing (to ensure compliance with the SLA) cloud monitoring is necessary. It provides information to cloud service providers to properly plan and manage various resources in order to deliver assured QoS to meet SLA parameters. Both commercial and open source cloud monitoring platforms are available which provide detailed, timely and accurate monitoring information which is further analyzed for resource provisioning and trouble shooting. The monitoring metrics and tests are classified as computation based and network based. Computation based metrics are server throughput, CPU -speed, time per execution, utilization, memory pages exchanged per unit time, memory/ disk throughput, VM startup time, VM acquisition and release time etc. whereas network based monitoring metrics are round trip time, jitter, network throughput, traffic volume, network bandwidth etc. To monitor and manage complex cloud infrastructure, the monitoring system must be fast, efficient, scalable, effective, robust, adapt easily to dynamic behavior of requests and workloads.

D. Parameters for Load Balancing

There are number of parameters on the basis of load is examined on a particular node to determine if it has pre-requisite load and it should not be overutilzed or underutilized. In this paper four parameters have been taken for detailed literature review.

1. Execution time based on CPU utilization- The overutilization and underutilization of hosts can be identified by monitoring their CPU utilizations. If the utilization increase some maximum threshold or appears below minimum threshold, their VM consolidation should be carried out to optimize the performance of the nodes.

2. Service Level Agreement Violations (SLAV) - CPU is the main cloud resource such that its overutilization may result in SLA violations. SLA Violations (SLAV) refers to the event when VM cannot get the requested amount of Millions of Instructions per Second(MIPS) which may occur when VM share the same host with other VMs cannot get the required CPU performance due to consolidation [13].

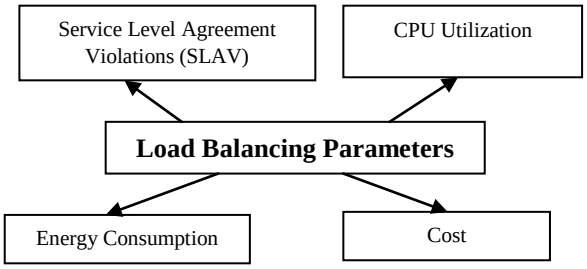


Fig.3. Load Balancing Parameters

3. Energy Consumption- Datacenters consume power to remain operational. This energy requirement is so enormous that on average a single data center consumes energy equivalent to 25,000 households [13]. Green Grid and Green cloud computing are the innovations to ensure optimal utilization of resources and reduce the energy consumption. To measure carbon emissions from datacenters, carbon usage effectiveness (CUE) metric is proposed [12] which is the ratio of total carbon emissions by a datacenter to the power consumed by it. The overall power utilized by a datacenter depends on energy requirements of CPU, memory, Disk, routers and other network devices. So, VMs of underutilized hosts must be migrated to other hosts and then turning them off to save electricity. In this context, load balancing is done so that there are no underutilized nodes.

4. Cost- Researchers have tried to minimize the cost and optimize the allocation of tasks each processor. Cost factor for task scheduling and load balancing on multiple processors in cloud environment can be analyzed from different perspectives like execution (operating) cost and delay cost. Operating cost is the total of cost for executing all jobs (total load) on available processors Delay Cost is defined as the total delay for all jobs waiting for allocation to processors and hence for execution. The delay occurs due to time lapse between the submission of a task, its allocation to a processor and further its completion of execution on that processor. Sometimes, if this delay is above the acceptable deadline time of the task then the task can be considered failed. Total cost for allocating and executing jobs on a single

node is the sum of operating cost and delay cost. Further, this cost for all the processors is added up to get the overall cost for executing jobs on multiple processors in cloud.

In this paper, a review of various scheduling and load balancing algorithms using optimization techniques based on SLAV, CPU utilization, energy consumption and operating cost has been done.

III. Related Work

Extensive literature is available on scheduling independent tasks or workflows while efficiently distributing load among VMs hosted on physical machines. Review has been done based on the parameters mentioned in previous section.

A. Time and Cost

There is trade-off between the time taken to complete the execution of tasks and the operating cost incurred. In other words, lesser cost is incurred if the tasks (may be dependent or independent) have flexible time deadlines within which they must be executed.

Rajiv, Liang et al. [2] proposed a decentralized, scalable, fault tolerant peer-to-peer cloud resource provisioning and application management technique. They applied extended Distributed hash tables (DHT) to search a node for VM deployment. DHT use efficient indexing and query routing techniques to handle multi-dimensional queries of cloud environment. The performance has been evaluated on the basis of response time, coordination delay with network complexity. Their results show both response time and coordination delay increased with increase of problem complexity since the task threads have to wait for longer period of time due to increased problem size.

Soni et al. [10] addressed workload as a multi-objective problem comprising of number of tasks where each task must be finished within its deadline time. Authors have taken three objectives-to balance Load, maintaining request priority and minimizing the execution error. A new optimization technique is presented based on bee colony optimization algorithm for scheduling the load for efficient utilization of resources with minimum latency. The fitness function minimizes the difference between the requested load and the served load with minimum execution error. Users are assigned priority on the basis of their payment for cloud resource usage such that free users are given least priority. One who paid more is given high priority and marked as primary users while free users as secondary users. The results are compared with Artificial Bee colony Single objective, FIFO and Random algorithm. Load is scheduled while maintaining the priorities among requests and reducing the number of unscheduled tasks under heavy load conditions.

Monir and Mohammed [3] used Divisible Load theory (DLT) to divide the load into fractions and scheduling them among available processors to minimize the overall total cost. DLT is used as a tool to handle extensive computation load in areas of aerospace data, image processing, non-linear processing and scheduling large scale divisible data in Grids and in cloud computing environment. However, this will achieve load balancing as processing nodes will receive fraction of load and no node will be overloaded. DLT calculates the optimal fraction of load that is to be allocated to each processor. Authors claim to achieve decrease in delay cost and hence the total cost is minimized.

The problem of average relative percentage of imbalance (ARPI) minimization is introduced by T. Keskinturk et al. [4] with sequence-dependent setup times. Their proposed mathematical model implements improved versions of both ant colony metaheuristic optimization algorithm (ACO*) and Genetic Algorithm (GA*). They also used nine heuristics for load balancing in parallel machines environment, and shows that ACO* gives better results than nine heuristics and GA*.Improvements in ACO* and GA* results in better performance by using two exchange heuristics for local search problem. Their results of ACO* algorithm show much lower ARPI as compared to all heuristics and ACO, GA, GA* metaheuristics.

Hybrid optimization based on Metaheuristic Swarm Optimization (MSO) and Cuckoo Search (CS) is used in [5] for task scheduling in cloud environment. Their scheduling model has three modules- scheduler system, set of cloudlets and VMs, mapping algorithm which assign a cloudlet to a particular VM which take minimum time for allocation and execution. The objective is to minimize the makespan time, response time and to maximize the resource utilization. In this hybrid approach, authors have used MSO as main search algorithm for scheduling cloudlets (tasks) on VMs while CS is used to speed up the convergence by improving the search space used by MSO.

The increased complexity of workflow applications in heterogeneous environment like cloud can be tackled using hybrid techniques to solve the workflow based scheduling problem. Another scheduling strategy with load balancing of VMs on hosts based on Genetic Algorithm is presented in [6] by Jianhua et al. The schedule with least influence on the current system status after resources are mapped to VMs is considered as the best solution to achieve load balancing and best resource utilization with minimum dynamic VM migrations.

B. Service Level Agreement Violations

Buyya et al. [7] presents a high level architecture in the form of a working prototype to evaluate the performance on the basis of feasibility and effectiveness of SLA-based resource provisioning and management in cloud environment. Various challenges to SLA-based resource provisioning have been discussed with further exploration for SLA oriented resource allocation techniques is suggested to enhance the efficiency of system, minimize SLA violations and to increase the overall profit for cloud service provider.

Cloud resource provisioning architecture which incorporates load balancing is given by Stefano et al. [8]

which effectively satisfies QoS requirements for application requests by users and optimize the resource utilization. SLA violations are controlled by load balancer to distribute the request load across multiple resources and to ensure QoS. In case SLA is violated, then predefined time constraint is provided within which the QoS is maintained otherwise the service provider face penalty. The parameters mainly focused are performance based SLA policies- response time and system availability. When SLA violation is detected, the application request is mapped onto a new VM and when the VM is no longer required, it is de-allocated to optimize the resource usage.

An SLA based novel heuristic for scheduling is proposed by Vincent et al. in [9] to minimize the cost and maximize the resource utilization based on amount of CPU cycles, network bandwidth, and storage required for application deployment in cloud platform. It also includes a load balancer to uniformly distribute the load among VMs on physical hosts with automatic activation of new VM when available VMs are insufficient to handle the incoming load. Their results show improved scheduler performance and deciding the optimal resource for the incoming request based on multi-objective performance parameters.

C. Energy Efficiency

To increase the service provider profit and hence, the overall performance, Jing Liu et al. [1] have designed a Multi-objective Genetic Algorithm (MO-GA) based on reduced the power consumption model. The objective is to minimize energy consumption and maximize service provider profit while following the deadline constraint. Cloud users requests heterogeneous cloud resources and these requests are dynamic in nature. The service request is decomposed into different levels of granularity with varying CPU utilizations. Their results of simulation are compared with Maximum Applications scheduling algorithm and random scheduling algorithm using three task arrival rates- low, medium and high. Their results show that at low arrival rate MO-GA achieves good results of 44.46% less energy consumption with 5.73% profit gain.

The usage of renewable energy sources (Solar, Wind, Hydro) results in decreasing the carbon footprint and hence the effect of green house gases but, the main challenges towards their utilization to power data centers are that these sources of energy are inconsistent, irregular, unpredictable and uncertain in nature. Taxonomy of recent research and challenges in managing and applying renewable energy in data center is provided by W. Deng et al. in [12]. When, where and why to migrate from grid to renewable energy to power up data centers to result energy efficiency and minimum operational cost. Depending upon the cost of grid energy and availability of renewable energy, datacenters can be powered up by either of these energy sources. When the excess renewable energy is available, this can be stored onto battery backup for future use

Green Cloud Architecture given by Anton B. et al. [13] considered CPU utilization as the overall system load which shows that the CPU utilization varies with time as workload fluctuates. Energy efficient architectural principles, resource allocation policies with scheduling algorithms preserving QoS have been addressed in this paper. Idle hosts are either switched off or set to sleep mode. Authors follow Minimization of Migration (MM) Policy in case CPU utilization crossed upper threshold and the Highest Potential Growth (HPG) Policy. Their results shows that under dynamic workload scenarios, the cloud computing model delivers highly improved cost effective and energy efficient performance.

Moona Yahchi et.al [14] presented Cuckoo Optimization Algorithm (COA) based approach to detect overloaded hosts and migrates their VMs to underutilized hosts using Minimum Migration Time (MMT) policy without resulting more overloaded hosts. Further, VMs of underutilized hosts are migrated to other hosts, turning them to sleep mode to reduce power consumption and hence achieving energy efficiency. They claimed to achieve about 121 times less energy consumption than non-power aware policy.

The utilization patterns of hosts are used to decide the number of hosts that remain active as in [15]. For this information, the utilization state hosts is determined using Double –Threshold Energy Aware Load Balancing (DT-PALB) algorithm taking two thresholds for CPU utilization- above 75% (over utilized node) and below 25% (underutilized node). Their results achieved reduction of 26.41% energy consumption as compare standard PALB algorithm by reducing the number of power-on machines. J. Doyle et al. [10] observed the challenge of receiving dynamic requests for cloud services across the globe, proposed the Stratus system based on Vornoi partitions. Their model details and minimizes the carbon emissions (carbon intensity- g/KWh), electricity cost, cooling cost and average job time taken for execution. Authors claim to achieve 21% reduction in carbon emission, 61% electricity save and average best effort time taken reduced by 47% as compare to Round Robin baseline. Beloglazov and Rajkumar Buyya [24] evaluate various heuristics for dynamic VMs reallocation based on live migration and current CPU requirements for CPU performance. Their proposed technique saves extensive energy while ensuring QoS parameters. Depending on the appropriate number of operating node, a power aware load balancing (PALB) algorithm for IAAS heterogeneous architecture is proposed by Jaffery et al. [11] which claims to decrease the power consumption of cloud datacenter by 70%-97% by considering the utilization patterns of the resources at a time.

Table I shows the analysis of the previous work done for resource allocation (scheduling) and load balancing in order to improve the utilization of cloud resources. The comparison has been done on the basis of four parameter mentioned in section 2.

Table 1. Comparison of various Load Balancing Techniques based on Time, Cost, SLA violations and Energy Efficiency

Parameter	Author	Problem undertaken	Optimization technique applied	Nature of Tasks / Inputs	Goals Achieved	Simulation environment
Time and Cost	Rajiv, Liang et al. [2]	cloud resource provisioning, application management and Load Balancing	Decentralized peer-to-peer technique based on distributed Hash Tables for efficient query routing and discovery	Independent Applications	Improved performance decreased response time, delay and increased Scalability	Aneka using Amazon Machine Images (AMIs)
	Monir et al. [3]	job scheduling	Divisible Load theory(DLT) model for scheduling divisible load	Workflows	minimize the overall processing time for scheduling jobs cost benefit for provider	CloudSim
	T. Keskinturk et al. [4]	Parallel Machine Scheduling and Load Balancing	Ant colony optimization algorithm and genetic algorithm	Independent Jobs taken randomly	minimize average relative percentage of imbalance (ARPI) ACO outperforms GA for Local Search	NA
	P. Durgadevi et al. [5]	VM placement and task scheduling	Amalgamation of Swarm Optimization (SO) algorithm and Cuckoo search (CS) algorithm	Workflows	reduction of the makespan time with increased resource utilization	CloudSim
	Jianhua Gu et al. [6]	Scheduling VMs on physical Machines with Load Balancing	Genetic Algorithm using minimum cost for generating schedule as optimization criteria.	VM List for server mapping	Better Load Balancing and resource Utilization minimum VM migration	CloudSim
SLA Violations	Buyya et al. [7]	SLA-oriented dynamic resource allocation and management	Architecture integrating market based provisioning policies with virtualization technologies	CPU-intensive application	Meet QoS requirements using Minimum resources while meeting Deadlines	Aneka with Amazon EC2 (in real time)
	Stefano et al. [8]	QoS based SLA Monitoring and resource utilization optimization	Dynamic reconfigure based resource allocation to avoid reaching max threshold	Dynamic workload	95% SLA efficiency with minimum VM allocations	Discrete-event based simulator, having different software component
	Vincent et al. [9]	Novel heuristic for Scheduling and Load balancing using multiple SLA parameters	Dynamic resource allocation and provisioning based on fixed resource and on-demand resource.	heterogeneous workloads based on Web and HPC Applications	On-demand resource allocation show better results in terms of resource utilization and deployment efficiency	CloudSim
	J. Doyle et al. [10]	Reducing carbon emissions, operational Cost and average service request time	Vornoi partitions based Stratus system to find optimal path from request to datacenter depending on -geographical distance, government Policies and carbon emission price	Independent Applications Workload	Balancing carbon emissions, operational cost and average service request time while conforming to priorities and SLA	Self developed simulation environment
	Jaffery et al. [11]	SLA based dynamic configuration, management and optimization of resources in cloud computing	middleware architecture with load balancer to ensure QoS otherwise dynamically reconfigure additional resources	Dynamic workload	Optimized resource utilization without SLA violations	Self developed simulation environment

Energy Efficiency	Jing Liu et al. [1]	task scheduling model	multi-objective genetic algorithm (MO-GA)	Independent Applications	increased service providers profit and reduced power consumption for low arrival rate	CloudSim
	W. Deng et al. [12]	Challenges of volatile, unpredictable variable, irregular nature of renewable energy	renewable energy generation models with capacity planning of data centers for scheduling and load balancing	NA	Decreased Energy consumption by data centers when cloud providers know when, where and how to leverage renewable energy in datacenters	NA
	Anton B. et al. [13]	challenges for Architectural Principles, resource allocation policies, QoS based scheduling algorithms for energy-efficient cloud computing	An architectural framework based on MM policy and principles for energy-efficiency in Cloud computing are defined and evaluated	Independent Applications Workload	Reduced energy consumption in Cloud data centers based on dynamic reallocation of VMs based on current CPU utilization	CloudSim
	Moona Yahchi et.al [14]	Detection of over-utilized and migration of VMs to underutilized hosts for load balancing	Overutilized hosts are detected using Cuckoo optimization Algorithm (COA) and their VMs are migrated using Minimum Migration Time (MMT) policy to the other hosts.	Independent Applications Workload	Reduction in power Consumption and hence, reducing CO_2 production while least SLA violations as compare to other power aware models	CloudSim
	J. Adhikari et al. [15]	Energy consumption in cloud computing with load balancing	Double Threshold Energy Aware Load Balancing (DT PALB) algorithm to track the computing nodes state and determining the number of nodes that should be active for current utilization patterns	Independent Applications Workload	Reduced power consumption by minimizing the number of power-on machines as compare to Round Robin scheduling algorithm	CloudSim

IV. Conclusion

Resource provisioning and Load Balancing are two important paradigms in cloud computing that effect the resource utilization, time to service the requests (including response time, delays, makespan time, execution time) and energy consumed by nodes in datacenters while ensuring the QoS requirements as expected by cloud customers. These parameters must be achieved without violating the SLA between the customer and the cloud service provider. In this paper, an extensive literature survey has been done for resource allocation to application requests (may be independent or workflow based) along with balancing the workload among the VMs to ensure no node is underutilized or overutilized. Further, a comparative study based on different parameters has been done to understand various optimization techniques used for scheduling and load balancing that will provide as future scope to study further for budding researchers.

References

[1] J. Liu, X. Luo, X. Zhang, F. Zhang and B. Li, "Job Scheduling Model for Cloud Computing Based on Multi-Objective Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 3, January 2013.

[2] R. Ranjan, L. Zhao, X. Wu, A. Liu, A. Quiroz and M. Parashar, &ldquo,Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing& rdquo, Cloud Computing, Computer Communications and Networks, N. Antonopoulos and L. Gillam, eds., pp. 195-217, Springer, 2010.

[3] M. Abdullah and M.Othman, "Cost-based Multi-QoS Job Scheduling Using Divisible Load Theory in Cloud Computing", Procedia Computer Science, vol. 18, pp. 928-935, 2013.

[4] T. Keskinturk, M. Yildirim and M. Barut, "An ant colony optimization algorithm for load balancing in parallel machines with sequence-dependent setup times", Computers & Operations Research, vol. 39, no. 6, pp. 1225-1235, 2012.

[5] P. Durgadevi and Dr. S. Srinivasan, "Task Scheduling using Amalgamation of Metaheuristics Swarm Optimization Algorithm and Cuckoo Search in Cloud Computing Environment", Journal for Research, vol.1,no. 9, 2015.

[6] GU, J. HU, T. ZHAO and G. SUN, "A New Resource Scheduling Strategy Based on Genetic Algorithm in Cloud Computing Environment", JCP, vol. 7, no. 1, pp. 42-52,2012.

[7] R. Buyya , S.K. Garg and R.N. Calheiros, "SLA-Oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions", Proc. International IEEE Conf. Cloud and Service Computing (CSC), pp. 1-10, 2011.

[8] Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, "QoS-Aware Clouds," in Proc. IEEE 3rd Intern. Conf. on Cloud Computing (CLOUD'10), pp. 321-328, 2012.

[9] Vincent C. Emeakaroha, Ivona Brandic, Michael Maurer, Ivan Breskovic, "SLA-Aware Application Deployment and Resource Allocation in Clouds", 35th IEEE Annual Computer Software and Application Conference Workshops, pp. 298-303,2011

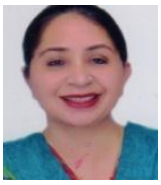
[10] J. Doyle, R. Shorten and D. O'Mahony, "Stratus: Load Balancing the Cloud for Carbon Emissions Control", IEEE Transactions on Cloud Computing, vol. 1, no. 1, pp. 1-13, 2013.

[11] Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky. "Power Aware Load Balancing for Cloud Computing",

Proceedings of the World Congress on Engineering and Computer Science 2011 Vol. IWCECS 2011, October 19-21, San Francisco, USA.
[12] W. Deng, F. Liu, H. Jin, B. Li and D. Li, "Harnessing renewable energy in cloud datacenters: opportunities and challenges", IEEE Network, vol. 28, no. 1, pp. 48-55, 2014.
[13] A. Beloglazov, J. Abawajy and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", Future Generation Computer Systems, vol. 28, no. 5, pp. 755-768, 2012.
[14] M. Yakhchi, S. Ghafari and S. Yakhchi, "Proposing a load balancing method based on Cuckoo Optimization Algorithm for energy management in cloud computing infrastructures", in 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), Istanbul, 2015.
[15] J. Adhikari and S. Patil, "Double threshold energy aware load balancing in cloud computing", in Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference, Tiruchengode, 2013, pp. 1 - 6.
[16] Joshi, Manveer and Bikram Pal Kaur., "Coap Protocol For Constrained Networks", International Journal of Wireless and Microwave Technologies, vol. 5, no. 6, pp. 1-10, 2015.

Authors Profiles

Amanpreet Kaur is an Assistant Professor in Dept. of Information Technology at Chandigarh Engineering College, Landran (Mohali). She completed her B.Tech in Computer Science and Engineering from Guru Nanak Dev University, Amritsar in year 2000 with distinction and honours. She received her M.Tech degree in Information Technology from Guru Nanak Dev University, Amritsar in year 2005 and topped in the University. She has been in teaching profession for the last 12 years and pursuing Ph.D. in Computer Science from IK Gujral Punjab Technical University, Jalandhar in the area of Cloud Computing. Her areas of interest are cloud computing, Operating Systems and Advanced Computer Architecture.



Dr. Bikrampal Kaur is Professor in the Dept. Of Information Technology at Chandigarh Engineering College, Landran, Mohali She holds the degrees of B.Tech. M.Tech, and M.Phil. She completed her Ph.D.in the field of Information Systems from Punjabi University, Patiala. She has more than17 years of teaching experience and served many academic institutions. She is an Active Researcher who has supervised many B.Tech. Projects ,MCA/M.Tech. dissertations and guiding Ph.D. research scholars . She contributed more than 40 research papers in various national & international conferences. Her areas of interest are Information System, ERP

Dr. Dheerendra Singh, having B.E., M.Tech, PhD in Computer Science & Engineering, is working as Associate Professor in Dept. of Computer Science & Engineering at Chandigarh College of Engg. And Technology, Chandigarh. He is life Member of IETE, New Delhi (Member No. M- 208777). He has published and presented more than 35 research papers in National & International Journals /Conferences. He is having 15 years of experience of teaching at various reputed Engineering Institutes which includes 7 years of experience as Head of Department. He is guiding Ph.D. and M.Tech students in Computer Science & Engineering.

How to cite this paper: Amanpreet Kaur, Bikrampal Kaur, Dheerendra Singh,"Optimization Techniques for Resource Provisioning and Load Balancing in Cloud Environment: A Review", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.9, No.1, pp.28-35, 2017. DOI: 10.5815/ijieeb.2017.01.04

