ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

A SURVEY ON RESOURCE SCHEDULING IN CLOUD COMPUTING

[1]C.Vijaya[*], [2]DR.P.Srinivasan
Research Scholar, Vellore Institute of Technology, Vellore, Tamilnadu, India-632 014.
Associate Professor, Vellore Institute of Technology, Vellore, Tamilnadu, India-632 014.

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Cloud computing is the Internet based approach to scale the resources up or down as required dynamically, popular for its elasticity property. In general, scheduling is done at 2 levels,during tasks to virtual machine allocation and virtual machine to physical machine allocation. The consumers request the resources based on their requirement of their applications. If there is a need for an application to use the resource aggressively but only for particular period of time (seasonal usage of resources) then we can provision the resources from the cloud on pay -per-usage basis. In data centers when the request for the resources are very large, then a scheduling technique for provisioning the resources from the physical machines to the virtual machines according to the consumer's requirement must be implemented so as to increase the performance and availability of the resource and to minimize the cost of using it. Both the providers and consumers must be benefitted by the resources on cloud environment. By utilizing the resources optimally, power consumed can also be reduced thereby saving the revenue for both the providers and consumers of the cloud. Scheduling is done based on energy consumption, bandwidth utilization, following SLA, virtualization techniques, cost optimization and load balancing. The survey lists the various scheduling algorithms used to schedule the resources as per the user demand and their performance in datacenters.

Keywords: Bandwidth utilization, Cost optimization, Datacenter, Provisioning, SLA, Virtualization,

Introduction

In this smart world, Cloud Computing grabs the attention of researchers, businessmen and many large organizations as it applies a technology which virtualizes the resources including the hardware, software, database and bandwidth etc., thereby saving the cost of using the resources for both the providers and consumers of the service. It minimizes the complexity of building the infrastructure, purchasing the resources, licensing the software in the client's perspective and

increases the revenue for the service providers who own the resources. According to NIST definition, "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storages, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[1]. Cloud Computing is a coarse grained approach based on utility computing. We can utilize the services of the cloud as we use electricity, water, gas facility in our homes and pay for the resources only for what we have used. Refer Fig I. It can provision hardware, software and datasets dynamically on demand to the customers in the datacenters. A datacenter is the collection of physical machines consisting of all the resources where the service is provided by the owners called providers.

The physical machine can have multiple virtual machines on which the tasks are allocated. Virtual machine is the instance of the server. Virtual machines are created with the help of Hypervisor or VMM(Virtual Machine Monitor). Hypervisor is software which is installed over the operating system to create multiple virtual machines with different configurations on which the user tasks are made to run. Examples: VMware, xen etc.

Basics of Cloud Computing

Though cloud computing grows rapidly, there are some challenges which are to be dealt with.

The challenges [1] of cloud computing which are common today are:

Security: As connectivity and sharing increases, security is becoming more important. which acts as the stopping force of advancement of cloud computing and its successor technologies like Fog computing and Internet of Things.

Efficient load balancing and scheduling: For the effective utilization of resources, the load from the clients has to be distributed among the servers evenly.

Performance monitoring consistent and robust service abstractions: Software which provides to the service provider the entire status of the cloud resources monitoring the services of the cloud.

Scale and Qos Management: Elasticity is one of the important features of IAAS model. The resources can be scaled up or down as per the need at the time of execution of user's applications. Sometimes there may arise a busty need for the resources. For example, any web service at a particular point of time may be used by millions of users and it gets suppressed after a particular duration of time like watching a cricket match.

Requires a fast and speed internet connection: The bandwidth of the network should be utilized properly such that the communication of the services should be fast. The bandwidth should be divided into channels to be used to its full capacity. Most of the times, the total bandwidth is not utilized. When it is not used so, the power gets wasted causing carbon dioxide emission, causing global warming.

2.1 Characteristics of the cloud

There are essentially five characteristics of the cloud which are called as the key attributes.

Shared Pool Resources: The resources are retrieved from a common set of resources as the common set provides economies of scale and high efficiency. Heterogeneous resources are collected to provide the service.

Broad network Access: IP (Internet protocol, HTTP (Hyper Text Transfer Protocol) and ReST(Representational State transfer) Protocols are used for network access. Resources are used from anywhere with an internet connection.

On-demand self service: The services are completely automated. The delivery of the service is in few seconds. Users can abstract from the implementation.

Scalable and elastic: The resources can be scaled up or down based on the need at that point of time.

Metered by use: The resource usage is monitored and managed for the service providers and consumers. Billing is done based on the usage of resources by the clients on pay-per use approach.

2.2 Cloud deployment

Cloud hosting deployment models explains the exact category of cloud environment which are distinguished by proprietorship, size and access. It defines the purpose and nature of the cloud.

Public: it is a type of cloud hosting in which the cloud services are delivered over internet which is open for a large public owned by an organization selling cloud services. The clients do not have the control over the infrastructure. The security offered to the public cloud is lesser when compared to the private cloud. Public cloud is better suited for business requirements which require managing the load. Due to the decreased overhead and operational cost, this is economical. The cost is shared among all the users of the cloud.

Examples: Google APP Engine, Microsoft Windows Azure,IBM Softcloud,Amazon EC2.

Private: Private cloud is also called as internal cloud. A cloud which belongs to a particular organization has been implemented which is safeguarded using a firewall. It permits only authorized users and gives the organization direct

control over the data. Business that have unexpected needs, mission-critical are better suited to adopt private cloud. Private cloud may exist off premises and managed by a third party.

Examples:

Eucalyptus, Ubuntu Enterprise cloud, Amazon Virtual private cloud

VM ware cloud infrastructure suite, Microsoft ECE datacenter

Hybrid: A type of cloud computing which integrates both public and private model is hybrid model. An arrangement of two or more cloud server that is bound together but remain individual entities. It permits the user to increase the capability by aggregation or customization with another cloud package. It adapts between two clouds, private and public as per the demands of the user. Business that has more focus on security and demand for their unique business can implement hybrid cloud as an effective business strategy. When extra demand arises for the resource, it is acquired for a particular application from the public cloud. This process is called **cloud bursting**.

Community: A type of cloud hosting in which the setup is manually shared between many organizations that belong to a particular community like banks and trading firms. It is multitenant setup which is shared among many organizations that belong to a specific group which has similar computing applications. The members share similar security, performance and privacy concerns. A community cloud is appropriate for organizations that work on joint ventures, tenders or researches that need cloud computing ability for managing, building and implementing similar projects.

2.3 Cloud architecture

There are three categories orlayers or **service models** that create cloud architecture. They are:

Software As AService (SAAS): Also called as on-demand software. Software is offered as a service to the clients as a web based service. This allows organization to access business functionality at a very lower cost when compared to paying for licensed applications and using hardware for that application. Users gain access to software and database. The application instance or the application software and the database is shared by many tenants which is highly configurable and they use it as if the application is hosted on a dedicated server Cloud providers manage the infrastructure and platform that runs the application.

Platform As A Service (PAAS): PAAS is the outgrowth of SAAS application model. PAAS is a computing platform which includes Operating system, Programming language, execution environment, and database and web server. The

clients don't need to allocate resources manually. The service provider such as Microsoft Azure and Google App Engine scale automatically to match the applications demand. PAAS refers to capability provided to consumer to deploy consumer created or acquired applications on to the cloud infrastructure. Users can develop web application without installing any tools and deploy those applications without any system administration skills. It includes services for concurrency management, fault tolerance and security. Also integrates with

web providers to offer a development environment to the client services and databases.

Infrastructure As A Service (IAAS): It offers Platform virtualization as a service. Virtualized software and servers are offered to the customers. IAAS is hosted in cloud data centers. Clients don't have to purchase any software, hardware or licensing. It is the capability given to the consumers to provision the resources. Resources include processing, storage, networking and other fundamental computing resources. In the provisioned resource the consumer is able to deploy applications and have the limited control of selected network, then deploy and run the applications. The customers can provision the resource from service providers and can rent to others in IAAS. He has control over the deployed applications; he can create his own application or purchase from third party. Service level agreements are not violated. Billing is done automatically based on utility computing. Refer Figure II.

re are three major components [2] which forms the cloud architecture.

Client: A system using the services of another system is called a client. End-user, tenant, consumer are the different names of the client who uses the services of the cloud.

Server: The servers are the system which provides access to the clients.

Datacenter: Datacenter is the collection of servers hosting variety of applications and services. It may be located at large distance from the client's location.

Scheduling

Scheduling is vital and it is done at places where operational computing is required as in virtual machines and in physical machines Scheduling on cloud computing is done at two levels. One is scheduling the tasks to the virtual machines and provisioning the resources on the physical machine to the virtual machines (also called as images). The resources should be at provisioned to the virtual machines by using hypervisors, matching the user's requests. Scheduling is based on various metrics like power consumption, SLA, virtualization, cost, load balancing etc.

There are two types of scheduling algorithms namely static and dynamic. For conventional computing, static algorithm is chosen where the request for CPU is received and made to wait in the queue. After a period of time, all the requests are allocated to the corresponding resources. But in dynamic scheduling which is obvious in cloud computing paradigm, the requests are heterogeneous in nature meaning that each request to the server may follow a different data format, data structure, instruction set etc.,

Static scheduling method: The client requests for the resources needed for his application to run on it and then he submits the job to the resource. Here the resources may be underutilized or over utilized during the time of the execution. A schedule which runs too many jobs on the physical machine is overloaded and little number of jobs is under loaded. For example, The Round Robin algorithm which is a static algorithm. it allocates the job to the resource only for a particular period of time slice and preempts the tasks. The tasks are then moved to the waiting queue and give a chance to the next job.

Dynamic scheduling: The user submits his job and when the application is running on the cloud, the resources are allocated as per the need of the application and so the instances of the application are created without provisioning extra resource to the application. So the wastage of the resource is avoided in the datacenters. Whatever size of the resource needed alone is configured by the virtual machine and so overutilization is not done. When the resource is overloaded, redistribution of resources should be done to prevent overloading. For example: FCFS algorithm provisions the resources to the virtual machines depending on the arrival time of the tasks as it comes on the fly. There are many meta-heuristic scheduling algorithms existing like ACO (Ant Colony Optimization) algorithm and its variations PCO (Particle swarm optimization algorithm and its variations and Honey Bee Algorithm etc.The other dimensions of scheduling are classified as:

a) **Resource scheduling:** The scheduling of resources (Physical machines) to the virtual machines comes under resource scheduling or resource provisioning.

b) **VM scheduling:** Every application request from the user to the cloud service needs heterogeneous requirement of resources, So the virtual machine images (instances) are allocated according to the needs of the application.

c) **Workflow scheduling;** This type of scheduling optimizes the execution time of the workflow of the job. It maps the dependent tasks on parallel resources for execution.

d) **Task scheduling:** A particular task is selected and serviced by the virtual machine. The VMs are managed by provisioning, de-provisioning and migration of processes between the physical machines.

The management of provisioning and releasing the resources is an important issue to be solved. The resources are provided as the service which is the functionality of the resources. Cloud is the connection of all the resources that are idle which are available throughout the world and provisioning the resources to the needy clients on a rental basis. For doing so, proper allocation of tasks submitted by the clients should be done.

3.1 Resource Management

For scheduling the resources, Resource management is very important. It includes:

Resource discovery: Resource discovery is the process that identifies the list of all suitable physical resources on the cloud.

Resource scheduling: The scheduling is done at two stages. One is at job/task to virtual machine mapping and other one is at virtual machine to physical machine mapping. Resource scheduling is the process of selecting the best resource from the matched physical resource. It actually identifies the physical machine where the virtual machines are to be created to provision the resources from the cloud infrastructure. They objective of the scheduling is to use the resources to its full ability at low cost. It benefits both the provider and the user by earning more revenue to the provider side and saving the cost of using the resources to the client side. Job scheduling selects the next job to be executed on Virtual Machine.

Resource allocation: After the optimal resource has been identified, the process of allocating the selected resource to the job or task of user's request is done. Job is the collection of independent or dependent processes and each process is called a task. The job is submitted to the selected resource for providing best quality of service.

Resource monitoring: After the submission of the job to the selected resource, the resource needs to be monitored to avoid any mal functionality and to track the status of the job. The underutilization of the resources like hardware, memory and bandwidth is managed by a method called server consolidation.

Table-I

S.No	Technique/Algorithm implemented	Metrics improved /affected	Advantages
1	Berger Model [27]	Better fairness of resource allocation	Bandwidth is utilized effectivelyUser task is completed in minimum time

2	Online resource reservation scheme [28]	Improved Acceptance ratio and Reconfiguration cost	Bandwidth was scaled up and down dynamically Resource allocation at runtime Can expand the range of scheduling oppurtunities.
3	Ant colony algorithm PBACO [29]	Make-span, cost of user, deadline violation rate and resource utilization.	Improved completion time Better System performance(Load Balancing Cost of using the resources is decreased.
4	Container-Based Provisioning for DCP [30]	Reduce energy consumption and scheduling delay	Energy consumption is reduced. Scheduling delay is minimized Reconfiguration cost is also minimized.
5	"Skewness" Algortihm [31]	Reduced time and reduced no. of resources used (skewness metric)	Amount of servers in use is decreased Time Execution time is reduced.
7	PANDA, a framework for static scheduling BoT applications across resources in both private and public clouds[32]	best trade-off point between cost and performance	Application completion time is reduced Cost of using the resources is mimimized.
8	Greedy Scheduling, Round Robin Power Aware Best Fit Decreasing Algorithm[33]	Energy conservation is compared	Energy efficiency is analyzed in three algorithms and energy conserved.
9	ACO and VM DFS(Dynamic Forecast Scheduling) [34]	Performance and Cost	Lower resource wastage than VM_DFS algorithm Better Load Balancing among PMs
10	K-means Clustering Algorithm[35]	Energy cost of the cloud is reduced SLA violation decreased	Energy cost is reduced, SLA is not violated.
11	VM Placement Algorithm[36]	Guarantee the VM Performance	VM to PM mapping is effective in such a way as to not waste the resources like memory, Bandwidth along with the CPU.
12	New Map Reduce cloud service model, Cura for provisioning cost-effective Map Reduce services in a cloud [37]	Cost effectiveness and Performance improved	80 % reduction in the cost 65% reduction in response time with respect to face book like traces,
13	Simulated Annealing, Firefly Algorithm and Cuckoo Search Algorithm to find an optimal solution of the submitted tasks [38]	Reduced Processing time of VMs	Overall processing time of the VMs is reduced
14	Online Traffic Scheduling Algorithm[39]	On demand bandwidth and dynamic pricing	An effective Bandwith allocation Charging mechanism to reduce cost
15	Back Propagation Neural	Improved Economic	Predicts Price of the resources

	Network (BPNN) based price prediction algorithm and a Price Matching Algorithm[40]	efficiency	Determine transction efficiency.
16	Adaptive Multi-objective Task Scheduling (AMTS) Strategy[41]	Reduction in Processing time and transmission time	Reduces Energy consumption , Increases QoS Effective resource utilizastion
17	FCFS ,RR, Genetic Matchmaking[42]	Reduction in Execution time	Varying cost and computational efficiency are studied
18	Optimal Voltage Frequency Technique[43] DVFS (Dynamic Voltage Frequency Scaling) and ALR(Adapter Link Rate) techniques	BetterEnergy efficiency	Increases the energy efficiency and increases the performance of the resources.

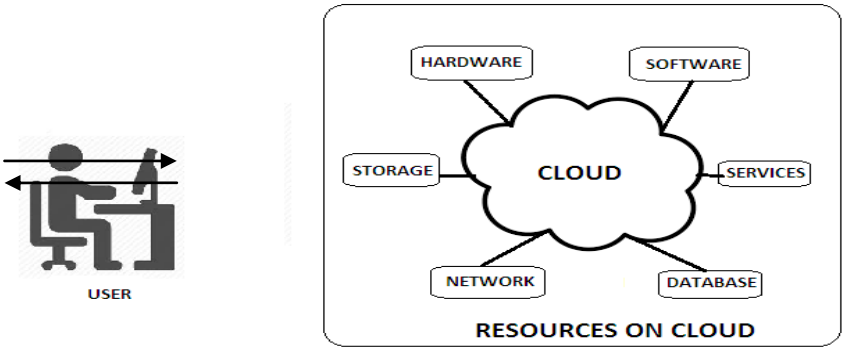

Fig.I: Resources on cloud utilized by the users.

SAAS(Software As A Service)
PAAS(PlatformAsA Service)
IAAS(Infrastructure As a Service)
RESOURCES
NETWORK BANDWIDTH
HARDWARE

Fig.I: The hierarchy of the services over the resources on Internet.

Literature Survey

The scheduling of the cloud is classified based on Energy efficiency, SLA, Virtualization, Cost optimization and Load Balancing metrics.The job should be scheduled to the virtual machines considering the above metrics. Power

consumption should be reduced in order to save money and also to protect the earth from the emission of carbon dioxide which is dealt separately as green computing. The SLA between the provider and the consumer should be followed to achieve reliability of the cloud. Cost has to be optimized for both the provider and consumer in such a way that both of them must be benefitted. Virtualization should be done to avoid overloading or under loading of the machines utilized thereby saving cost and power. Server consolidation is a technique used to achieve overloading or under loading. Load balancing is the process of allocating the job equally to all the servers so that the jobs can be finished in time avoiding overutilization or underutilization of resources. Several papers have been studied and analyzed regarding scheduling of resources in the cloud environment.The scheduling of the cloud is classified based on Energy efficiency, SLA, Virtualization, Cost optimization and Load Balancing metrics. After the job is scheduled to the virtual machines we need to optimize multiple metrics as we saw above. Power consumption should be reduced in order to save money and also to protect the earth from the emission of carbon dioxide which is dealt separately as green computing. The SLA between the provider and the consumer should be followed to achieve reliability of the cloud. Cost must be optimized for both the provider and consumer getting benefits. Server consolidation is done to avoid overloading or under loading of the machines utilized thereby saving cost and power. Load balancing is the process of allocating the job equally to all the servers so that the jobs can be finished in time avoiding overutilization or underutilization of resources. Several papers have been studied and analyzed regarding scheduling of resources in the cloud environment based on different metrics.

The dynamic creation and placement of virtual machines[6] has been discussed by Vikash et Al. The VM placement on physical machine is created by three classes namely Reservation, on-demand and Spot market. The VM placement algorithm maps the VM onto available PM in the datacenter so as to increase the performance of the server and to minimize the power consumption. Rank scheduling algorithm is used by Open Nebula VMware whereas Greedy First Fit Scheduling Algorithm and Round Robin algorithm are used by both Eucalyptus and Nimbus VMware. Experiments on Task scheduling were conducted by placing VMs on the PMs using timeshared and space shared policies by the use of Cloudsim tool.

Renu Bala et Al. have compared different scheduling algorithms for cloud Scheduling. Heterogeneous Earliest Finish Time (HEFT) algorithm is a scheduling algorithm which schedules the application to the heterogeneous processor. But it is a static algorithm and not suitable for dynamic request of resources. Resource Aware Scheduling Algorithm (RASA) is

a combined approach of two algorithms namely, Max-Min and Min-Min algorithms. Min-Min executes smaller tasks first and then bigger tasks. When smaller tasks are more than the bigger tasks, we use Max-Min algorithm which schedule bigger tasks first and then smaller tasks. Scalable Heterogeneous Earliest Finish Time (SHEFT) algorithm is the algorithm where the resources can be elastically scaled but cannot be reserved based on the workflow size. Here the scheduling is done in three phases. They Are Task prioritizing, task selection and processor selection phase. So, two steps are followed here. One is HEFT and the other is resource allocation mechanism so as to minimize the execution time. In improved cost based algorithm for task scheduling, the tasks areranked as high priority, low priority and medium priorityand critical tasks are sent to the specialized processor for those tasks, where the other tasks are executed on other processors.

In [8], different types of virtualization namely Application virtualization, Desktop virtualization, Network virtualization, server virtualization, presentation virtualization and management virtualization have been discussed. The data center capacity to deploy applications demanded by the user has been analyzed. The analysis is done using cloudsim , a simulator tool and Cloud Analyst tool based on High Performance computing network. The code for simulating a datacenter using cloudsim tool has also been given.

Allocation of the virtual machines to the physical machine according to the user's need improves the response time [9]. Response time is taken as a metric to be improved in this paper. An allocation method which allocates the web server to the virtual machines has been proposed. In this system architecture, a client submits the job to be executed on the cloud server. A job scheduler schedules the job to the VM scheduler after multifactor (CPU, time, storage etc.,) verification. The VM scheduler does the migration of the process to the available resource. If the selected resource is overloaded, then the process is migrated to some other virtual machine on some other node (Physical machine).A resource optimization framework has been proposed and implemented on dot net platform. The result is, the waiting time and response time are minimized in turn reducing the memory space.

Neeraj mangal et al.[10] has given a brief classification of scheduling based on energy conservation, SLA, virtualization and cost effectiveness metrics. Various algorithms regarding different kinds of scheduling have been discussed and a detailed comparison of these algorithms with performance metrics and results has been listed. The challenges in cloud computing like providers profitability, customer satisfaction are identified. The power consumed by the datacenters has

been minimized using many algorithms which have been listed. In service level agreement scheduling, both the providers and tenants have to agree on certain conditions to satisfy the quality of service and reliability. Violation of SLA is a key issue in this paper. To avoid the deadline miss rate which is a SLA metric, an algorithm has been developed to control the admission of the users which is considered as a general approach for managing the cloud service. In virtualization based scheduling, papers regarding migration of processes, server consolidation, provisioning of VM, de-provisioning of VM, mapping VMs to PMs have been studied and analyzed. All the scheduling proves to improve the quality of service and make-span. In cost optimization kind of algorithms, the cost of using the resources has been decreased by implementing several techniques. A survey of algorithms used to conserve power has been studied in [11]. The power consumption in datacenters leads to excessive operating costs and carbon dioxide emission according to this paper. Power consumption is classified into two categories viz., static power consumption and dynamic power consumption. Static power consumption is the power utilized by the system components. Dynamic power consumption is the power utilized during the operation of the system components. The power is managed through many algorithms like VM consolidation algorithm. Ant Colony Optimization, SNOOZE etc. A variety of algorithms which are used for reducing power consumption in data centers have been analyzed.

An adaptive resource scheduling strategy [12] which calculates the running time of the job and allocates the job to the virtual machines has been discussed. The logic behind is, it optimizes the number of virtual machines used for each job, and also increases the user reliability proposed by Guisheng Fan et Al. A resource scheduling model has been proposed in which the reflection mechanism is used where relationship of jobs and deadline of jobs were considered as the metrics.CTL (Computation Tree Logic) has been used to describe the properties of resource scheduling model. An adaptive resource scheduling algorithm has been proposed which schedules the user request dynamically, re-optimize and redistribute the resources dynamically.

The energy consumed has been reduced through dynamic capacity provisioning, Harmony, a heterogeneous aware dynamic capacity provisioning framework for cloud datacenters[13].K-means clustering algorithm has been implemented to divide the workloads (user requests) into different groups based on resource and performance requirements. It is proved that the DCP (Dynamic Capacity Provisioning) algorithm consumes 28% less energy than any other algorithm.DCP algorithm can be applied to the public clouds. Workloads are heterogeneous with respect to the size

of the job on private clouds, where the tenants can choose the virtual machine size.DCP provides a solution in such kind of situations. Many heterogeneous servers of different manufacturer were taken into consideration to form a cluster and the tasks were allocated to the servers so as to minimize the power consumption using Harmony.

A characterization study of how virtual machines are used in datacenters on a private cloud, virtual machines are consolidated to form a bigger machine and when the migration of virtual machines are done has been analyzed [14].Virtual Machine deployment is sharing the physical resource between the virtual machines implemented on it. VM deployment may be static or dynamic. Migration can be done after switching off the virtual machine and this process is described as static deployment or also can be done while the virtual machine is running without changing the resource requirement to some other physical node and this process is described as dynamic or live deployment.VM resource provisioning is allocating the virtual resources to the VM according to the task requirements. According to this paper, migration is done for server consolidation and scalable resource provisioning. The frequency of resource provisioning is analyzed in real cloud environment. It has been proved that most of the virtual machines seldom migrate. A general view of how virtualization technology is applied in the datacenters has been shown through experiments The results show that the virtual machines are always on.

The VM placement problem in a physical machine containing more than one core CPU has been addressed by Zoltán et Al[15]. A physical machine or virtual machine may have many cores of CPU. When a VM is mapped to a PM, then each core of the vCPU must be mapped to the core of the pCPU. This paper ignores the physical core during the VM placement which leads to the suboptimal VM placement. The hypervisor (software used to create VM) does the core scheduling (allocation of vCPUs to pCPUs).The total CPU load of VMs should not exceed that of the total capacity of PMs. A framework has been implemented called CLPFD (Constant Logic programming Finite Domains) and two-stage CP algorithm has been implemented. It checks if the vCPUs of the VM in the pCPU currently and the vCPUs of the new VM might be mapped to Physical machine.. Also it checks the core to where the VM should be placed. It has been proved that the core level placement information was important to obtain good performance and this algorithm implements the reduction of cost.

D.Sireesha et al. proposes a new system architecture for designing a VM scheduler which consists of a Predictor, Hot spot and Cold spot solver[16]. Based on the heuristics, the predictor predicts the expected load on each physical machine

in future and the resource demand for the VMs. Hot spot solver checks if any Physical machine is over utilized (i.e.) beyond its threshold. If yes, some of VMs are migrated to other physical machines. Cold spot solver checks if the power consumption is above the threshold, then that PM is turned off to control the power consumption. "Skewness" algorithm has been proposed to control the unevenness in the utilization of multiple resources on a server. Based upon thedemand of the application, mapping VMs to PMs is done adaptively.

Min-Min algorithm calculates the minimum execution time of the job and the jobs are sorted according to the minimum execution time from which the job containing the minimum value of the minimum execution time is assigned to the processor first and this is followed for other jobs[17].Max-Min is same as above but the maximum value of the minimum execution time is assigned to the processor first and this is followed for other jobs.Load balancing algorithm called RASA (Resource Aware Scheduling algorithm) has been proposed. This also uses the Min-Min and Max-Min strategy to start its execution. The expected response time of each virtual machine is provided to the scheduler by returning their ids to the scheduler. The scheduler finds which one is having lesser response time andhaving lesser no of jobs on it, and is given priority to assign the new request. Once the request is completed, the scheduler de-allocates by running de-allocation algorithm and the table is updated. The experiment is simulated on Cloudsim tool increasing the performance and decreasing the response time.

Lyapunov optimization [18] is an approach which maximize the time-average profit in the inter datacenters connected across the world. Here an algorithm to schedule the multicast oriented tasks in inter-datacenter networks has been proposed thereby increasing the time - average profit. Earlier works studied in the paper introduced a similar algorithm but in intra data centers.

The bandwidth of the network is utilized to the peak [19] only during certain period of time and all the other times, it is not used to its fullest capacity thereby just wasting the power consumption and results in wastage of energy. The bandwidth allocation problem is a big issue which was analyzed in this paper. Ting Wang et al. have proposed 2 schemes to solve this bandwidth problem, Firstly a multi commodity flow problem is taken to find the set of optimal routes in order to mitigate the power consumption. The NP hardness(due to high no. of nodes and routes) is calculated to run the algorithm and the NP hardness problem is solved using Artificial intelligence technique called Blocking Island scheme based heuristic (BHS) . a bandwidth allocation scheme has been designed which allocates the bandwidth effectively

using BHS. A Topology scheme based heuristic (THS) algorithm is also designed to calculate the minimum network subset and to maintain the traffic in the subset. The BHS and THS approaches are implemented on a simulator called DCNSim simulator. It has been proved that bandwidth and power are used effectively than the traditional use of power and bandwidth in datacenters. Snehanshu Saha et Al have shown the different operational costs for the network usage in datacenters and the costs for using resources available in the data centers[20]. Server and power cooling cost is a great issue. These two metrics are taken into account for conducting the experiments. They divide the expenditure into operational expenditure which includes power/cooling cost, maintenance cost etc whereas the capital expenditure is new server cost, infrastructure cost etc. Cobb-Douglas production function is implemented to make the inputs (cost of servers, networking, infrastructure and power), for which the maximum revenue for the input does not exceed a particular amount. A resource cost model has been proposed which defines the demand of tasks to resources [21]. Make-span (Execution time) and the user's budget are taken as metrics. Performance and cost are the multi-objective factors of this paper. Deadline violation rate and resource utilization is also considered as metric in some experiments conducted by the same authors. An improved Ant Colony Optimization algorithm was proposed and a better performance and cost metrics have been achieved. Ant Colony optimization problem is used to solve combinatorial optimization problem. The solution is based on Ant's behavior of secreting pheromone on finding its way from its home to the destination where food is available. Other Ants follow the pheromone smell in the path and reach the destination. They follow different path and the one which is optimal is followed by most of the ants. This nature of the ants is used in this ACO algorithm. This is used to schedule without considering the budget constraint. Algorithm was compared against FCFS and Min-Min algorithm and proved to be better than those two in terms of performance and cost.

For integration of enterprise systems, workflow scheduling is used. In [22], execution time and cost is reduced using trust service oriented workflow scheduling algorithm. This algorithm uses direct trust and recommendation trust to calculate trust metric.ARS algorithm was proposed o dynamically compute adaptive scheme for jobs based on attributes defined by the user. The ARS algorithm has been compared with FIFO, Load Balancing scheduling and Fairness job scheduling and proved that the algorithm works well for reducing the execution time and cost of the resources.

In big data, the query is submitted to the server which is very large in processing the data. So the query processing is split into 3 phases and provided to the different resources for processing that too in a heterogeneous way [23]. For doing so,

the resources should be allocated through an energy efficient algorithm. A map reduce framework called PRISM , a resource aware scheduler for Hadoop Map Reduce consists of a job scheduler which minimizes job completion time and maximizes resource utilization. The run time has been drastically reduced when compared to Hadoop. This architecture is implemented with a phase level scheduling algorithm on Hadoop 0.20.2.*

After proper resource provisioning, the VMs and all the resources should be monitored to diagnose any errors [24]. Different kinds of cloud monitoring platforms have been discussed like monitis, logic monitor, cloud watch, Open nebula, ANEKA, NEW Relic, etc. These are evaluated by using metrics such as performance; cost etc. The cloud monitoring tools gathers the information regarding network and the system. Monitoring is done to check the SLA violation or to migrate the VMs etc., Network monitoring is done on centralized architecture and decentralized architecture. Single point of failure, scalability issues are being monitored in centralized architecture. A comparison chart has been provided to show monitoring on various parameters. In data centers, it is found that the consolidation of servers reduces no. of storagedevices [25]. The task description is very important in task scheduling for a cluster of heterogeneous multi-core servers. The proposed Energy efficient extended load balancing algorithm has been implemented. Additional servers are included in the storage unit to implement this algorithm. The algorithm sends the VMs to sleep mode when it is idle or heavily used. Then the task is assigned to the virtual servers using FCFS approach for the remaining virtual servers. The algorithm has been given in the paper. It is proved that there is an improvement in the response time, segregation of storage space and energy conservation.

Conclusion

In this paper, we have discussed the scheduling strategies followed in the cloud environment. Different research works based on resource scheduling were analyzed. Scheduling is done between tasks and resources. Allocation of resources to virtual machines called resource provisioning, virtual machines to physical machines consisting of single core and multiple cores were analyzed. After scheduling has been done, monitoring was also done to follow up the resources. Using scheduling techniques, performance, bandwidth utilization and quality of service could be increased with the decrease in make-span time and cost. Virtualization is the technology in cloud computing which makes the resources spread across the world available to the end-users on a marginal cost. The various available scheduling algorithms in cloud computing environment and the performance which are improved or affected were listed out. More research work

needs to be done on virtualization, where provisioning of resources need attention to decrease the execution time and to improve availability of resources.

References

1. Palak Shrivastava Sudheer Kumar Arya , Dr. Priyanka Tripathi "Various Issues & Challenges of Load Balancing Over Cloud: A Survey" in International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issues 8 Aug 2016, Page No. 17517-17524.
2. Lipsa Tripathy, Rasmi Ranjan Patra "Scheduling In Cloud Computing" in International Journal on Cloud Computing: Services and Architecture (IJCCSA) ,Vol. 4, No. 5, October 2014 DOI : 10.5121/ijccsa.2014.4503
3. Jose M.AlearazCalcro, Senior IEEE, Juan Gutierrez Aguado "MonPaaS: An Adaptive Monitoring Platform as service for cloud commuting infrastructures and services" Citation Information 10,1109,TSC 2014 2302810, IEEE Transactions on Services Computing.
4. S.Sujan and R.Kanniga Devi " A Batchmode Dynamic Scheduling Scheme For Cloud Computing "Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015).
5. SaurabhBilgaiyan, SantwanaSagnika, MadhabanandaDas "Workflow Scheduling in Cloud ComputingEnvironment Using Cat Swarm Optimization" , IEEE Transactions On Parallel And Distributed Systems, Vol. 26, No. 5, May 2015.
6. Vikash ,Rao and Pahalad Singh "Dynamic Creation and Placement of Virtual Machine Using CloudSim ",International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014).
7. Renu Bala1, Gagandeep Singh2,Sahil Vashist , " A Review of Cloud Based Schedulers on cloud computing environment" in International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 7 July, 2014 Page No. 6866-6870.
8. AnkushDhiman, Mauli Joshi, "Analysis of Performance for Data Center under for Private Cloud through Cloud Computing" in International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 6 June, 2014 Page No. 6422-6431

9. Dr. D .Ravindran , Mr. S.Lakshmanan, "Performance Enhancement System for the Cloud with Multi Factor Resource Allocation Technique" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 5 Issue 8 August 2016 Page No. 17627-17632
10. Neeraj Mangla, Manpreet Singh, Sanjeev Kumar Rana, " Resource Scheduling In Cloud Environmet: A Survey" in Advances in Science and Technology Research Journal ,Volume 10, No. 30, June 2016, pages 38–50 DOI: 10.12913/22998624/62746.
11. Auday Al-Dulaimy, Wassim Itani, Ahmed Zekri1, and Rached Zantout, "Power management in virtualized data centers: state of the art", Journal of Cloud Computing: Advances, Systems and Applications (2016) 5:6 DOI 10.1186/s13677-016-0055-y.
12. Guisheng Fan, Huiqun Yu, Senior Member, IEEE, and Liqiong Chen , " A Formal Aspect Oriented Method for Modeling and Analyzing Adaptive Resource Scheduling in Cloud Computing", IEEE Transactions On Network And Service Management, Vol. 13, No. 2, June 2016.
13. Qi Zhang, Student Member, IEEE, Mohamed FatenZhani, Member, IEEE, RaoufBoutaba, Fellow, IEEE, and Joseph L. Heller stein, Fellow, IEEE, "Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud, IEEE Transactions On Cloud Computing, Vol. 2, No. 1, January-March 2014.
14. Robert Birke_, Andrej Podzimeky, Lydia Y. Chen_, and EvgeniaSmirniz, " Virtualization in the Private Cloud: State of the Practice", IEEE Transactions on Network and Service Management, , Citation information: DOI 10.1109/TNSM.2016.2601646.
15. ZoltánÁdámMann, " Multicore-aware virtual machine placement in cloud data centers", IEEE Transactions on Computers, Citation information: DOI 10.1109/TC.2016.2529629.
16. D.Sireesha, M.Chiranjeevi, P.Nirupama, "Distribution of cloud resources dynamically by using virtualization In International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 7 July, 2014 Page No. 7380-7383.
17. S. MohanaPriya, B. Subramani , "A New Approach For Load Balancing In Cloud Computing", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 5 May, 2013 Page No. 1636-1640

18. Kaiyue Wu, Ping Lu, and Zuqing Zhu, Senior Member, IEEE, "Distributed Onlin Scheduling and Routing of Multicast-Oriented Tasks for Profit-Driven Cloud Computing" IEEE Communications Letters, VOL. 20, NO. 4, APRIL 2016.
19. Ting Wang, Bo Qin, Zhiyang Su, Yu Xia, Mounir Hamdi1, Sebti Foufou and Ridha Hamila, "Towards bandwidth guaranteed energy efficient data center networking" in Journal of Cloud Computing: Advances, Systems and Applications (2015) DOI 10.1186/s13677-015-0035-7.
20. Snehanshu Saha, Jyotirmoy Sarkar, Avantika Dwivedi, Nandita Dwivedi, Anand M. Narasimhamurthy and Ranjan Roy "A novel revenue optimization model to address the operation and maintenance cost of a data center" in Journal of Cloud Computing: Advances, Systems and Applications (2016) 5:1 DOI 10.1186/s13677-015-0050-8
21. Liyun Zuo, Lei Shu, (Member, IEEE), Shoubin Dong, Chunsheng Zhu, (Student Member, IEEE), and Takahiro Hara, (Senior Member, IEEE), "A Multi-Objective Optimization Scheduling Method Based on the Ant ColonyAlgorithm in Cloud Computing", date of publication December 17, 2015,date of current version December 23, 2015.Digital Object Identifier 10.1109/ACCESS.2015.250894017524.
22. WenAn Tan, Yong Sun, Ling Xia Li, GuangZhen Lu, and Tong Wang , " A Trust Service Oriented Scheduling Model for Workflow Applications in Cloud Computing" in IEEE SYSTEMS JOURNAL, VOL. 8, NO. 3, SEPTEMBER 2014.
23. M.Sneha Priya, Mrs.R.Rebekha, " Heterogeneous Phase-Level Scheduling With Jobs Execution Scheduling Algorithm To Enhance Job Execution And Resource Management In Mapreduce" in International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issues 6 June 2016, Page No. 16880-16885.
24. Deepika Upadhyay, Sanjay Silakari, Uday Chourasia, "A Comprehensive Analysis of Cloud Computing Including Security Issues and Overview of Monitoring" in International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume - 3 Issue -9 September, 20Citation information: DOI 10.1109/TNSM.2016.2601646, IEEE
25. N.Susila and Dr.S.Chandramathi, "Energy Efficient Extended FCFS Load Balancing In Data Centers of Cloud" in International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016)
26. Baomin Xu, Chunyan Zhao, Enzhao Hu,Bin Hu, "Job scheduling algorithm based on Berger model in cloud environment " in Advances in engineering software, Volume 42, Issue 7, July 2011, Pages 419–425.

27. Carlofuerst, Stefan schmid,lalith suresh,paolo costa,"Kraken: Online and elastic resource reservations for multi-tenant datacenters" in IEEE infocom 2016- the 35th international conference on computer ommunications, Resource Reservations for Multi-tenant Datacenters" in EEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications , **DOI:** 10.1109/INFOCOM.2016.7524466.
28. Liyun Zuo, Lei Shu, Shoubin Dong, "A Multi-Objective Optimization Scheduling Method Based on the Ant Colony Algorithm in Cloud Computing" in IEEE Acess special section on Big data services and computational intelligence for industrial systems.DOI: 10.1109/access.2015.2508940.
29. Qi Zhang, David. R, "Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud" in IEEE Transactions On Cloud Computing, Vol. 2, No. 1, January-March 2014. DOI: 10.1109/Tcc.2014.2306427
30. NARAYANA RAO Appini, DEEPIKA.Tenepalli, "Active Resource Provision in Cloud Computing Through Virtualization" in IEEE international conference on Computational Intelligence and Computing Research 2014.
31. Mohammad Reza Hoseiny Farahabady, Young Choon Lee, and Albert Y. Zomaya, Fellow, IEEE, "Pareto-Optimal Cloud Bursting" in IEEE Transactions on Parallel and Distributed Systems (Volume: 25, Issue: 10, Oct. 2014).
32. Goyal, Seema Bawa, Bhupinder Singh, Sudhir , "Experimental Comparison of Three Scheduling Algorithms for Energy Efficiency in Cloud Computing" inCloud Computing in Emerging Markets (CCEM), 2014 IEEE International Conference ,DOI: 10.1109/CCEM.2014.7015491
33. Milad Seddigh, Hassan Taheri, Saeed Sharifian, "Dynamic prediction scheduling for virtual machine placement via Ant Colony Optimization" in Signal Processing and Intelligent Systems Conference (SPIS), 2015.
34. Qingxin Xia, Yuqing Lan, Liang Zhao, Limin Xiao, "Energy-saving Analysis of Cloud Workload Based on K-means Clustering" in Computing, Communications and IT Applications Conference (ComComAp), 2014 IEEE
35. Hui Zhao , Qinghua Zheng , Weizhan Zhang, "Virtual Machine Placement Based on the VM Performance Models in Cloud" in Computing and Communications Conference (IPCCC), 2015 IEEE 34th International Performance,DOI: 10.1109/PCCC.2015.7410296
36. BalajiPalanisamy, Member, IEEE, Aameek Singh, Member, IEEE, and Ling Liu, Senior Member, IEEE , "Cost-Effective Resource Provisioning for MapReduce in a Cloud" in IEEE Transactions on parallel and Distributed Systems, Vol.26, No.5, May 2015.

37. Tripti Mandal and Siyankar Acharyya, "Optimal Task Scheduling in Cloud Computing Environment: Meta Heuristic Approaches", in Proceedings of International Conference on Electrical Information and Communication Technology (EICT 2015)
38. Weijie Shi, Chuan Wu , Zongpeng Li, "A Shapley-value Mechanism for Bandwidth On Demand between Datacenters" in IEEE Transactions on Cloud Computing (Volume: PP, Issue: 99),DOI: 10.1109/TCC.2015.2481432,Date of Publication: 23 September 2015.
39. Xingwei Wang, Xueyi Wang, HaoChe, Senior Member, IEEE, Keqin Li, Fellow, IEEE,Min Huang, and Chengxi Gao, "An Intelligent Economic Approach for Dynamic Resource Allocation in Cloud Services" in IEEE Transactions on Cloud Computing (Volume: 3, Issue: 3, July-Sept. 1 2015).
40. Hua He ,Guangquan Xu ,Shanchen Pang "AMTS: Adaptive Multi-Objective Task Scheduling Strategy in Cloud Computing" in China Communications (Volume: 13, Issue: 4, April 2016),DOI: 10.1109/CC.2016.7464133.
41. Kavyasri M N, Dr. Ramesh B, "Comparative Study of Various Scheduling Algorithms in Cloud Computing" in International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 4 Issue 7 July 2015, Page No. 13364-13368. For Correspondence: