2013 Nirma University International Conference on Engineering (NUiCONE)

A Survey on Cloud Computing Resource Allocation Techniques

Swapnil M Parikh
Department of Computer Science and Engineering,
BITS edu campus, Varnama,
Vadodara, Gujarat, India
swapnil.parikh@gmail.com

***Abstract*-- Cloud Computing is a type of computing which can be considered as a new era of computing. Cloud can be considered as a rapidly emerging new paradigm for delivering computing as a utility. In cloud computing various cloud consumers demand variety of services as per their dynamically changing needs. So it is the job of cloud computing to avail all the demanded services to the cloud consumers. But due to the availability of finite resources it is very difficult for cloud providers to provide all the demanded services. From the cloud providers' perspective cloud resources must be allocated in a fair manner. So, it's a vital issue to meet cloud consumers' QoS requirements and satisfaction. This paper mainly addresses key performance issues, challenges and techniques for resource allocation in cloud computing. It also focuses on the key issues related to these existing resource allocation techniques and summarizes them.**

***Index Terms*-- Cloud Computing, Resource Allocation, Service Level Agreement, Virtualization.**

I. INTRODUCTION

Because of the advancement in Information and Communication Technology (ICT) over past few years, Computing has been considered as a utility like water, electricity, gas and telephony. These utilities are available at any time to the consumers based on their requirement. Consumers pay service providers based on their usage [2] [3].

Like all the other existing utilities, Computing utility is the basic computing service that meets the day to day needs of the general community. To deliver this vision, a number of computing paradigms have been proposed, of which the latest one is known as Cloud Computing. Cloud is nothing but large pool of easily accessible and usable virtual resources [2] [3].

Dr. Rajkumar Buyya says "A Cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers." [3]

Cloud computing is composed of three kind of services [1] [5] [6] [9].

1) Cloud Software as a Service (SaaS)

In this service model, instead of using locally run applications the cloud consumer uses the cloud provider's software services running on a cloud infrastructure. It is the job of cloud provider to maintain and manage the software services that are used by the cloud consumer. The cloud provider may charge according to quantity of software and using time. SaaS is the best way to use advanced technology. Salesforge.com and Customer Relationship Management (CRM) are the examples of such service model [1] [4] [5] [6] [7] [10].

2) Cloud Platform as a Service (PaaS)

In this service model, the cloud platform offers an environment on which developers create and deploy applications. It provides platform where applications and services can run. The consumers do not need to take care of underlying cloud infrastructure including network, servers, operating system or storage but has a control over deployed application. Google Application Engine, Microsoft Azure and RightScale are the example of such model [1] [4] [5] [6] [10].

3) Cloud Infrastructure as a Service (IaaS)

In this service model, cloud providers manage large set of computing resources such as storing and processing capability. Cloud consumer can control operating system; storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls). Sometimes it is also called as a **Hardware as a Service (HaaS)**. The cost of the Hardware can be greatly reduced here. Amazon Web Services, Open Stack, Eucalyptus, GoGrid and Flexiscale offers IaaS [1] [4] [5] [6] [10].

In cloud computing various deployment models have been adopted based on their variation in physical location and distribution. Regardless of the services, clouds can be classified among four models as mentioned below.

1) Private Cloud

It is private to the organization. All the cloud services are managed by the organization people themselves or any third party vendors as well as services are not provided to the general public. Private cloud may exist on premise or off premise [1] [5] [6] [8].

978-1-4799-0727-4/13/$31.00 ©2013 IEEE

2) Public or Hosted Cloud

All the cloud services managed by the organization are made available as in pay as you go manner to the general public. The business people can adopt such cloud to save their hardware and/or software cost. Public cloud may raise number of issues like data security, data management, performance, level of control etc [1] [5] [6] [8].

3) Community Cloud

Here cloud is available to specific group of people or community. All the cloud services are shared by all these community people. Community cloud may exist on premise or off premise [1] [5] [6].

4) Hybrid Cloud

It is a combination of two or more clouds (Private Cloud, Public Cloud, and Community Cloud) [1] [5] [6].

The rest of the paper is organized as follows: Section II presents issues and motivation related to resource allocation in cloud computing. Section III discusses various cloud computing resource allocation techniques proposed by researchers'. Section IV gives summary of all these existing resource allocation techniques with their used tools and possible improvements. Section V presents conclusion and discussion on resource allocation techniques.

II. Motivation

In cloud computing various cloud consumers demand variety of services as per their dynamically changing needs. So it is the job of cloud computing to avail all the demanded services to the cloud consumers. But due to the availability of finite resources it is very difficult for cloud providers to provide all the demanded services in time. From the cloud providers' perspective cloud resources must be allocated in a fair manner. So, it's a vital issue to meet cloud consumers' QoS requirements and satisfaction.

Traditional resource allocation techniques are not adequate for cloud computing as it is based on virtualization technology with distributed nature. Cloud computing introduces new challenges for manageable and flexible resource allocation due to heterogeneity in hardware capabilities, workload estimation and characteristics in order to meet Service Level Objectives of the cloud consumers' applications.

The ultimate goal of resource allocation in cloud computing is to maximize the profit for cloud providers and to minimize the cost for cloud consumers.

III. Literature Survey and Related Work

Qiang Li, Qinfen Hao, Limin Xiao and **Zhoujun Li** [11] proposed VM-base architecture for adaptive management of virtualized resources in cloud computing. Authors also designed a resource controller named Adaptive Manager that dynamically adjusts multiple virtualized resource utilization to achieve application Service Level Objective (SLO) using feedback control theory. Adaptive Manager is a multi-input, multi-output (MIMO) resource controller which controls CPU scheduler, memory manager and I/O manager based on feedback mechanism. To periodically measure application performance each Virtual Machine has sensor module which transmits information to the adaptive manager. Authors adopted Kernel based Virtual Machine (KVM) as a tool for infrastructure of virtual machine.

Mayank Mishra, Anwesha Das, Purushottam Kulkarni and **Anirudha Sahoo** [12] discussed that live virtual machine migration plays a vital role in dynamic resource management of cloud computing. Authors mainly focused on efficient resource utilization in non peak periods to minimize wastage of resources. In order to achieve goals like server consolidation, load balancing and hotspot mitigation, authors discussed three components – when to migrate, which VM to migrate and where to migrate – and approaches followed by different heuristics to apply migration techniques. Authors also discussed virtual machine migration over LAN and WAN with their challenges.

T. R. Gopalkrishnan Nair and **Vaidehi M** [13] presented a model, named as Ruled Based Resource Allocation (RBRAM) which deals with the efficient resource utilization in M-P-S (Memory-Processor-Storage) Matrix Model. Authors say that resource allocation rate should be greater than resource request rate. Major components of the system are: cloud priority manager, cloud resource allocation, virtualization system manager and end result collection. To analyse the performance of the cloud system authors considered the Cloud Efficiency Factor. However, authors also identified other parameters of Cloud System for future work.

Rosy Aoun, Elias A. Doumith and **Maurice Gagnairein** [14] proposed a model named as Mixed Integer Linear Program (MILP) for resource provisioning for enriched services in cloud environment. Authors stated that several basic services offered at IaaS level can be arranged together by the cloud providers for providing sophisticated services to the cloud consumers. Two original services, distributed data storage and multicast data transfer are jointly considered in addition to the traditional computing, centralized storage and point to point data transfer services. However, authors have considered the impact of four types of services: computing, storage, point to point data transfer and point to multipoint data transfer. The numerical results were given by considering 18-node backbone network.

Justin Y. Shi, Moussa Taifi and **Abdallah Khreishah** [15] explored a simple quantitative Timing Model method for cloud resource planning. For the same they considered the estimated resource usage times in steady state. Authors had calculated Speed up for Parallel Resource Planning based on Parallel Matrix Multiplication. To investigate multiple important dimensions of a program's scalability, authors proposed quantitative application dependent instrumentation

method instead of qualitative performance models. Authors had mainly focused on application inter dependencies for cost effective processing.

Chenn-Jung Huang, Chih-TaiGuan, Heng-MingChen, Yu-WuWang, Shun-ChihChang, Ching-Yu Li and **Chuan-HsiangWeng** [16] proposed resource allocation mechanism based on Support Vector Regression (SVR) and Genetic Algorithm (GA). Authors designed Application service prediction module with Support Vector Regression (SVR) to estimate the number of resource utilization according to the Service Level Agreement (SLA) of each process. Then authors designed global resource allocation module with Genetic Algorithm (GA) to redistribute the resources to the cloud consumers.

Zhen Xiao, Weijia Song and **Qi Chen** [17] aimed to achieve two goals – overload avoidance and green computing - for dynamic resource allocation through virtualization technologies. Based on dynamically changing need of the cloud consumers the designed and implemented system multiplexes virtual to physical resources adaptively. The multiplexing is done through Usher Framework. Authors designed a load prediction algorithm to predict future resource utilization without seeing into virtual machines. Authors had used "skewness" metric to measure uneven utilization of server. For the same they defined concept of "Hot Spots" and "Cold Spots" servers. In order to evaluate the performance of the algorithm designed authors used trace driven simulations.

Amit Nathani, Sanjay Chaudhary and **Gaurav Somani** [18] proposed an algorithm in a scheduler named Haizea for resource allocation policies like immediate, best effort, advanced reservation and deadline sensitive. Haizea is a resource lease manager that uses resource leases as resource allocation abstraction and implements these leases by allocating Virtual Machines (VMs). Authors main goal was to minimize resource rejection rate and reshuffle cost in order to provide all the above mentioned resource allocation policies for IaaS cloud. Authors also used two concepts named swapping and backfilling for deadline sensitive resource allocation policy. Authors mainly considered four lease parameters for their experiments: start time, duration, deadline and number of nodes.

Weiwei Lina, James Z. Wangb, Chen Liangc and **Deyu Qia** [19] proposed a threshold based dynamic resource allocation scheme for cloud computing. Authors mainly focused on application level resource allocation instead of mapping between physical resources and virtual resources for better utilization of resources. A threshold is used to optimize the decision of resource reallocation. The proposed algorithm consists of two procedures: **Datacenter**-resides at the datacenters central computer and **Broker**-runs on user's machine with the application. Both procedures interact with each other for dynamic resource allocation. The proposed algorithm is implemented by using CloudSim Toolkit.

Yichao Yang, Yanbo Zhou, Lei Liang, Dan He and **Zhili Sun** [20] focused on efficient data and network (combined) resource utilization for data intensive applications like IPTV. Authors proposed Cloud Infrastructure Service Framework (CISF) to achieve QoS requirements of cloud consumers. They introduced a Service-oriented Resource Broker (SRB) for guaranteed data transmission in cloud computing to discovery, select, reserve and assign data and network resources. Firstly the collected user requirements are given to Resource Requirement Interpreter to produce abstract resource requirement information. This information is then passed to the Resource Discovery Unit to produce list of resource combination which is passed to Resource Combination Ranker to assign priority. Finally Resource Reservation Unit makes coordinated resource reservation to resource gatekeepers through reservation interface.

Kejiang Ye, Xiaohong Jiang, Dawei Huang, Jianhai Chen and **Bei Wang** [21] proposed resource reservation based live migration framework of multiple virtual machines. The target machine in the framework holds four virtual machines: Migration Decision Maker, Migration Controller, Resource Reservation Controller and Resource Monitor. Authors focused on improving the migration efficiency through live migration of virtual machines and proposed three optimization methods: optimization in the source machine, parallel migration of multiple virtual machines and workload-aware migration strategy. To improve the migration efficiency authors had considered parameters like downtime, total migration time and workload performance overheads. Authors claimed that resource reservation strategy is required at source machine and target machine.

Congfeng Jiang, Xianghua Xu, Jilin Zhang, Yunfa Li and **Jian Wan** [22] raised the effective resource allocation problem based on real time knowledge of workload and performance feedback of running services. Authors had proposed stochastic model of resources in virtualized environments. Authors had also proposed resource allocation and scheduling heuristics algorithms with service level agreement constraints. To improve the effectiveness of the future incoming dynamic workload, the performance of the targeted machine had been considered as a performance feedback mechanism to the source. This feedback mechanism improves the resource allocation method proposed by authors themselves.

GuiyiWei, Athanasios V. Vasilakos, Yao Zheng and **Naixue Xiong** [23] proposed game-theoretic method for fair resource allocation in cloud computing. Authors used Game Theory for QoS constrained resource allocation problem. Firstly, authors considered optimization problem for cloud services for which Binary Integer Programming method was proposed for initial optimization. Based on the initial result, an evolutionary mechanism was designed to achieve the final optimal and fair solution. In summary, authors focused on the sophisticated parallel computing problem on unrelated machines connected across the Internet.

Baomin Xu, Chunyan Zhao, Enzhao Hu and **Bin Hu** [24] proposed job scheduling algorithm based on Berger Model with dual fairness constraints. Authors had mainly concentrated on fairness of resource allocation and cloud consumers' satisfaction to the provided services. Based on parameters like completion time and bandwidth, cloud consumers' tasks had been classified. According to the characteristics and preferences of tasks, resources were assigned to the cloud consumers. Authors implemented their algorithm on CloudSim toolkit and compared with optimal completion time algorithm. Results show that algorithm based on Berger Model is better.

Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya [25] proposed resource allocation algorithm for SaaS providers who can minimize infrastructure costs and SLA violations and for SaaS consumers to assure service satisfaction. Authors had considered consumers Quality of Service parameters such as response time and infrastructure level parameters such as server initiation time. Authors implemented three cost driven algorithms from both consumers and SaaS providers perspective. The first algorithm is a base algorithm which maximizes the profit by minimizing the number of SLA violations. The second algorithm maximizes the profit by minimizing the cost by reusing VMs, which have maximum available space. The third algorithm maximizes the profit by minimizing the cost by reusing VMs, which have minimum available space. The second and third are proposed by authors which were simulated on CloudSim environment.

Borja Sotomayor, Ruben Santiago Montero Ignacio Martin Llorente and Ian Foster [26] presented a lease suspension/resumption time model for prediction of various run time overheads involved in using virtual machines through which advanced reservations can be made. Authors used Haizea, open source lease management architecture for advance reservation leases, best effort leases and immediate leases. As Haizea cannot operate on physical hardware resources, authors integrated Haizea with OpenNebula virtual infrastructure manager. Experiments were done on Xen Virtual Machine.

IV. Summary of Resource Allocation Techniques

Table 1 summarizes the work done by various researchers and future work and/or gaps in their existing work.

V. Conclusion and Discussion

Cloud Computing is the new era of computing for delivering computing as a resource. The success and beauty behind cloud computing is due to the cloud services provided with the cloud. Due to the availability of finite resources, it is very important for cloud providers to manage and assign all the resources in time to cloud consumers as their requirements are changing dynamically. So in this paper the problem of resource allocation with its different techniques in cloud computing environments has been considered.

Many authors have proposed algorithms and methods for dynamic resource allocation in cloud computing. In summary, an efficient Resource Allocation Technique should meet following criteria's: Quality of Service (QoS) aware utilization of resources, cost reduction and power reduction / energy reduction. Some of the authors have focused on IaaS based resource allocation with VM scheduling. The ultimate goal of resource allocation in cloud computing is to maximize the profit for cloud providers and to minimize the cost for cloud consumers.

TABLE I
Summary of Resource Allocation Techniques

Year	Author	Techniques/Algorithms	Tools used	Future work and/or gaps in existing technologies
2009	Qiang Li, Qinfen Hao, Limin Xiao and Zhoujun Li [11]	Adaptive Management of Virtualized Resources using Feedback Control	KVM	Only KVM model, network I/O performance, Still better modelling can be done for resource sharing.
2012	Mayank Mishra, Anwesha Das, Purushottam Kulkarni and Anirudha Sahoo [12]	Live Virtual Machine Migration	Not Mentioned	Only load on the virtual machine for migration is considered. Consumer requirements and priority of job is not considered.
2011	T. R. Gopalkrishnan Nair and Vaidehi M [13]	Rule Based Resource Allocation Model (RBRAM), M-P-S (Memory-Processor-Storage) Model	Not Mentioned	Only Cloud Efficiency Factor is considered to evaluate the performance level of cloud system.
2010	Rosy Aoun, Elias A. Doumith and Maurice Gagnairein [14]	Mixed Integer Linear Program for resource provisioning	Globus Toolkit and Condor dispatcher (Sebastian 2)	Dynamism is considered based on pre-planned traffic. Resource allocation algorithm is executed offline. Observed simulation runtimes exceed the considered scheduling period.
2011	Justin Y. Shi, Moussa Taifi and Abdallah Khreishah [15]	Timing Model with Amazon EC2	Amazon EC2	Authors mainly focused on cost effectiveness parallel processing
2013	Chenn-Jung Huang, Chih-TaiGuan, Heng-MingChen, Yu-WuWang, Shun-ChihChang, Ching-Yu Li and Chuan-HsiangWeng [16]	Resource Allocation based on Support Vector Regression (SVR) and Genetic Algorithm (GA)	CloudSim	Authors plan to modify the algorithms to decrease the calculation time in terms of the prediction process to improve the GA's convergence speed.
2012	Zhen Xiao, Weijia Song and Qi Chen [17]	Resource Allocation based on "Skewness" Metric	Xen Virtual Machine	The migration of job has to be there but which job to migrate that is not specified. The effect of partial machine migration is not discussed.
2011	Amit Nathani, Sanjay Chaudhary and Gaurav Somani [18]	Policy based resource allocation, Haizea scheduler	Haizea	Backfilling algorithm is not implemented yet. Response time of best effort service can be enhanced.
2011	Weiwei Lina, James Z. Wangb, Chen Liangc and Deyu Qia [19]	Threshold based dynamic resource allocation	CloudSim	Experiments were done only for Internet Applications. Overhead of virtual resources are considered not for physical resources.
2010	Yichao Yang, Yanbo Zhou, Lei Liang, Dan He and Zhili Sun [20]	Cloud Infrastructure Service Framework for QoS requirements with Service-oriented Resource Broker.	Not Mentioned	Optimization of combined resources is a challenge.
2011	Kejiang Ye, Xiaohong Jiang, Duwei Huang, Jianhai Chen and Bei Wang [21]	Live Migration of Virtual Machines	Xen and VMWare	Intelligent live migration machine can be future work.
2011	Congfeng Jiang, Xianghua Xu, Jilin Zhang, Yunfa Li and Jian Wan [22]	Resource allocation and scheduling heuristics algorithms based on real time knowledge of workload and performance feedback	Xen	Cache allocation and contention not considered which may improve prediction accuracy and allocation efficiency.
2010	GuiyiWei, Athanasios V. Vasilakos, Yao Zheng and Naixue Xiong [23]	Game theoretic method for fair resource allocation to solve sophisticated parallel computing problem	Not Mentioned	Applications are not clearly defined.
2011	Baomin Xu, Chunyan Zhao, Enzhao Hu and Bin Hu [24]	Job scheduling algorithm based on Berger Model based on dual fairness constraints	CloudSim	More accurate vector value of the general expectation vector can obtained.
2011	Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya [25]	Resource allocation algorithm for maximizing total profit and customer satisfaction.	CloudSim	Some more services can be included for improving the performance such as spot pricing.
2009	Borja Sotomayor, Ruben Santiago Montero Ignacio Martin Llorente and Ian Foster [26]	Prediction of various run time overheads for advanced reservations, Haizea based	Xen Virtual Machine	Experiments are done on only Xen VM, not on KVM VM or any other.