International Journal For Technological Research In Engineering
Vol. 1, Issue. 1, Sep – 2013 *ISSN (Online): 2347 - 4718*

A SURVEY ON SCHEDULING BASED RESOURCE ALLOCATION IN CLOUD COMPUTING

Shabnam Khan[1]
[1] Computer Science and Engineering Department,
Sobhasaria Engineering College,
Sikar, Rajasthan, India.
[1]khanshabnam07@gmail.com

Abstract: Cloud computing is the next generation of technology which unifies everything into one. It is an on demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. In this paper a review of various policies for dynamic resource allocation in cloud computing is shown based on Topology Aware Resource Allocation (TARA), Linear Scheduling Strategy for Resource Allocation and Dynamic Resource Allocation for Parallel Data Processing. Moreover, significance, advantages and limitations of using Resource Allocation in Cloud computing systems is also discussed. The other challenges of resource allocation are meeting customer demands and application requirements. In this paper, various resource allocation strategies and their challenges are discussed in detail. It is believed that this paper would benefit both cloud users and researchers in overcoming the challenges faced.

Keywords: Cloud Computing, Cloud Services, Resource Allocation, Infrastructure.

I. INTRODUCTION

Cloud computing emerges as a new computing paradigm which aims to provide reliable, customized and QoS (Quality of Service) guaranteed computing dynamic environments for end-users [22].Distributed processing, parallel processing and grid computing together emerged as cloud computing. The basic principle of cloud computing is that user data is not stored locally but is stored in the data center of internet. The companies which provide cloud computing service could manage and maintain the operation of these data centers. The users can access the stored data at any time by using Application Programming Interface (API) provided by cloud providers through any terminal equipment connected to the internet.

Not only are storage services provided but also hardware and software services are available to the general public and business markets. The services provided by service providers can be everything, from the infrastructure, platform or software resources. Each such service is respectively called Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service (SaaS) [45].

There are numerous advantages of cloud computing, the most basic ones being lower costs, re-provisioning o f resources and remote accessibility. Cloud computing lowers cost by avoiding the capital expenditure by the company in renting the physical infrastructure from a third party provider. Due to the flexible nature of cloud computing, we can quickly access more resources from cloud providers when we need to expand our business. The remote accessibility enables us to access the cloud services from anywhere at any time. To gain the maximum degree of the above mentioned benefits, the services offered in terms of resources should be allocated optimally to the applications running in the cloud. The following section discusses the significance of resource allocation.

Significance of Resource Allocation

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely.

a) **Resource contention** situation arises when two applications try to access the same resource at the same time.
b) **Scarcity of resources** arises when there are limited resources.
c) **Resource fragmentation** situation arises when the resources are isolated. There will be enough resources but not able to allocate to the needed application.
d) **Over-provisioning** of resources arises when the application gets surplus resources than the demanded one

Parameter	Provider	Customer
Provider Offerings	√	-
Resource Status	√	-
Available Resources	√	-
Application Requirements	-	√
Agreed Contract Between Customer and provider	√	√

Table. 1. INPUT PARAMETERS

II. RELATED WORK

Very little literature is available on this survey paper in cloud computing paradigm. Shikharesh et al. in paper [30] describes the resource allocation challenges in clouds from

the fundamental point of resource management. The paper has not addressed any specific resource allocation strategy. It is evident that the paper which analyzes various resource allocation strategies is not available so far. The proposed literature focuses on resource allocation strategies and its impacts on cloud users and cloud providers. It is believed that this survey would greatly benefit the cloud users and researchers.

III. RESOURCE ALLOCATION STRATEGIES (RAS) AT A GLANCE

The input parameters to RAS and the way of resource allocation vary based on the services, infrastructure and the nature of applications which demand resources. The schematic diagram depicts the classification of Resource Allocation Strategies (RAS) proposed in cloud paradigm. The following section discusses the RAS employed in cloud.

A. Execution Time

Different kinds of resource allocation mechanisms are proposed in cloud. In the work by Jiani at.al [15], actual task execution time and preemptable scheduling is considered for resource allocation. It overcomes the problem of resource contention and increases resource utilization by using different modes of renting computing capacities. But estimating the execution time for a job is a hard task for a user and errors are made very often [30]. But the VM model considered in [15] is heterogeneous and proposed for IaaS.

B. Policy

Since centralized user and resource management lacks in scalable management of users, resources and organization-level security policy [6], Dongwan et al. [6] has proposed a decentralized user and virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources. Based on role based access control (RBAC), virtualized resources are allocated to users through domain layer.

C. Virtual Machine (VM)

A system which can automatically scale its infrastructure resources is designed in [24]. The system composed of a virtual network of virtual machines capable of live migration across multi- domain physical infrastructure

IV. DIFFERENT RESOURCE ALLOCATION POLICIES

a. A time-driven adaptive mechanism for cloud resource allocation

Cloud computing service providers deliver their resources based on virtualization to satisfy the demands of users. In cloud computing, the amount of resources required can vary per user request. Therefore, the providers have to offer Different amounts of virtualized resources per request.To provide worldwide service, a provider may have data centers that are geographically distributed throughout the world. Likewise, the user locations vary in geographic location. Since cloud computing services are delivered over the internet, there may be undesirable response latency between the users and the data centers. Hence, for the best service, the provider needs to find a data center and physical machine that has a light workload and is geographically close to the user.The proposed model finds the best match for the user requests based on two evaluations: 1) the geographical distances between the user and data centers and 2) the workload of data centers. Hence, the model allows the users to find a data center that is guaranteed to be the closest distance and have the lightest workload. Also, it finds a light workload physical machine within the data center for a provider.

b. Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems

Resource allocation is the most important challenges in cloud computing. The service provider should work hard for allocating resources based on the client's SLA (Service Level Agreement).Force directed search algorithm is the solution for SLA based resource allocation problem for multi-tier applications in cloud computing .This algorithm considers the Gold SLA, and Bronze SLA. The provider gives the guarantee for the response time in Gold SLA. The requests are moved forward and backward in multi-tier service model. The server serves the backward requests. Probability Distribution Function (PDF) is used for finding the arrival rate in the Gold SLA. The resource management problem's aim is to maximize the total profit.

c. Multi-dimensional Resource Allocation Algorithm in cloud Computing

Cloud computing has emerged as a new technology and it has been increasingly adopted in many areas including science and engineering as well as business. How to arrange large-scale jobs submitted to cloud in order to optimize resource allocation and reduce cost is an issue of common concern. Paper present are two common ways to optimize resource utilization. One is at the application level when applications are arriving; other is in the period of applications running. In this paper, author makes effort on the former way to address multi-dimensional resource allocation problem by proposing a resource allocation scheme using fewer nodes to process user's applications. To address multi-dimensional resource allocation problem, raises several concerns.

d. SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments

SaaS is a software delivery method that provides access to software and its functions remotely as a Web-based service. It allows organizations to access business functionality at a cost typically less than paying for licensed applications since SaaS pricing is based on a monthly fee. In

order to deliver hosted services to customers, SaaS companies have to either maintain their own hardware or rent it from infrastructure providers. This requirement means that SaaS providers will incur extra costs. Though the cost of the resources has to be minimum, it is also important to satisfy a minimum service level to customers. Saas providers are able to manage the variety of customers, mapping customer requests to infrastructure level parameters and co nsider ing heterogeneity of Virtual Machines. The allocation method uses two different algorithms such as ProfminVmMaxAvaiSpace and ProfminVmMinAvaiSpace. First algorithm is designed to minimize the number of VMs by utilizing already initiated VMs. The criterion for reusing VM is, it should have maximum available space. The algorithm optimizes the profit by minimizing number of initiated VM. Moreover, it minimizes number of violations caused by service upgrade because VM has the maximum available space. In such a way, it reduces the penalty caused by upgrading service. The disadvantage of this algorithm is that it can decrease the profit. The maximum available space is occupied by small number of accounts and it leading other requests to be served by a new VM. To overcome the disadvantages of this algorithm, reducing the space wastage by using minimum available space (MinAvaiSpace) Strategy instead of MaxAvaiSpace Strategy. When there are more than one VM with same type, deployed with the same product type as customer request required, the VMs with enough available space to serve are selected. Then request is scheduled to the machine with the minimum available space in a best-fit manner). The proposed algorithms minimize the SaaS provider's cost and the number of SLAViolations based on the dynamic allocation of resources to requests.

e. Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems

In this paper authors propose an adaptive resource allocation algorithm for the cloud system with preempt able tasks in which algorithms adjust the resource allocation adaptively based on the updated of the actual task executions. Adaptive list scheduling (ALS) and adaptive min-min scheduling(AMMS) algorithms are used for task scheduling which includes static task scheduling, for static resource allocation, is generated offline. The online adaptive procedure is use for re- evaluating the remaining static resource allocation repeatedly with predefined frequency. In each reevaluation process, the schedulers are re-calculating the finish time of their respective submitted tasks, not the tasks that are assign to that cloud. Policy based resource allocation in IaaS cloud [6] Most of the Infrastructure as a Service (IaaS) clouds use simple resource allocation policies like immediate and best effort. Immediate allocation policy allocates the resources if available, otherwise the request is rejected. Best-effort policy also allocates the requested resources if available otherwise the request is placed in a FIFO queue. It is not possible for a cloud provider to satisfy all the requests due to finite resources at a time. Haizea is a resource lease manager that tries to address these issues by introducing complex resource allocation policies. Haizea uses resource leases as resource allocation abstraction and implements these leases by allocating Virtual Machines (VMs). Haizea supports four kinds of resource allocation policies: immediate, best effort, advanced reservation and deadline sensitive. Proposed dynamic planning based scheduling algorithm is implemented in Haizea that can admit new leases and prepare the schedule whenever a new lease can be accommodated. Experiments results show that it maximizes resource utilization and acceptance of leases compared to the existing algorithm of Haizea.

V. CONCLUSION

Scheduling is one of the most important tasks in cloud computing environment. In this paper, we have analyzed various scheduling algorithm and tabulated various parameter. We have noticed that disk space management is critical issue in virtual environment. Existing scheduling algorithm gives high throughput and cost effective but they do not consider reliability and availability. So we need algorithm that improves availability and reliability in cloud computing environment. In future enhancement will propose a new algorithm for resource scheduling and comparative with existing algorithms.

REFERENCES:

[1] A.Singh ,M.Korupolu and D.Mohapatra. Server-storage virtualization:Integration and Load balancing in data centers. In Proc.2008 ACM/IEEE conference on supercomputing (SC'08) pages 1-12, IEEE Press 2008.

[2] AndrzejKochut et al.: Desktop Workload Study with Implications forDesktop Cloud Resource Optimization, 978-1-4244-6534-7/10 2010 IEEE.

[3] Atsuo Inomata, TaikiMorikawa, Minoru Ikebe, Sk.Md. MizanurRahman: Proposal and Evaluation of Dynamin Resource Allocation Method Based on the Load Of VMs on IaaS(IEEE,2010),978-1-4244-8704-2/11.

[4] D. Gmach, J.RoliaandL.cherkasova, Satisfying service level objectives in a self-managing resource pool. In Proc. Third IEEE international conference on self-adaptive and self-organizing system. (SASO'09) pages 243-253.IEEE Press 2009.

[5] David Irwin, PrashantShenoy, Emmanuel Cecchet and Michael Zink:Resource Management in Data-Intensive Clouds: Opportunities and Challenges .This work is supported in part by NSF under grant number CNS-0834243.

[6] Dongwan Shin and HakanAkkan : Domain- based virtualized resource management in cloud computing.

[7] Dorian Minarolli and Bernd Freisleben: Uitlity –based Resource Allocations for virtual machines in cloud computing (IEEE, 2011), pp.410-417.

[8] DusitNiyato, Zhu Kun and Ping Wang: Cooperative Virtual Machine Management for Multi-Organization Cloud Computing Environment.

[9] FetahiWuhib and Rolf Stadler: Distributed monitoring and resource management for Large cloud environments (IEEE, 2011), pp.970-975.

[10] HadiGoudaezi and MassoudPedram: Multidimensional

SLA-based Resource Allocation for Multi-tier Cloud Computing Systems IEEE 4th International conference on Cloud computing 2011, pp.324-331.
[11] HadiGoudarzi and MassoudPedram: Maximizing Profit in Cloud Computing System Via Resource Allocation: IEEE 31st International Conference on Distributed Computing Systems Workshops 2011: pp, 1-6.
[12] Hien et al.,'Automatic virtual resource management for service hosting platforms, cloud'09, pp 1-8.
[13] Hien Nguyen et al.: SLA-aware Virtual Resource Management for Cloud Infrastructures: IEEE Ninth International Conference on Computer and Information Technology 2009, pp.357-362.
[14] I.Popovici et al,"Profitable services in an uncertain world". In proceedings of the conference on supercomputing CSC2005.
[15] Jiyani et al.: Adaptive resource allocation for preemptable jobs in cloud systems (IEEE, 2010), pp.31-36.
[16] Jose Orlando Melendez & shikhareshMajumdar: Matchmaking with Limited knowledge of Resources on Clouds and Grids.
[17] K.H Kim et al. Power-aware provisioning of cloud resources for real time services. In international workshop on Middlleware for grids and clouds and e-science, pages 1-6, 2009.
[18] Karthik Kumar et al.: Resource Allocation for real time tasks using cloud computing (IEEE, 2011), pp.
[19] Keahey et al., "sky Computing",Internet computing, IEEE,vol13,no.5,pp43-51,sept-Oct2009.
[20] Kuo-Chan Huang &Kuan-Po Lai: Processor Allocation policies for Reducing Resource fragmentation in Multi cluster Grid and Cloud Environments (IEEE, 2010), pp.971-976.
[21] Linlin Wu, Saurabh Kumar Garg and Raj kumarBuyya: SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments (IEEE, 2011), pp.195-204.
[22]Lizhewang,JieTao,KunzeM.,Castellanos,A.C,Kramer,D. ,Karl,w, "High Performance Computing and Communications",IEEE International Conference HPCC,2008,pp.825-830.
[23] M.SuhailRehman,Majd F.Sakr : Initial Findings for provisioning Variation in Cloud Computing(IEEE,2010),pp.473-479 .
[24] P.Ruth,J.Rhee, D.Xu, R.Kennell and S.Goasguen, "Autonomic Adaptation of virtual computational environments in a multi-domain infrastructure", IEEE International conference on Autonomic Computing, 2006,pp.5-14.
[25] Patricia Takako Endo et al.: Resource allocation for distributed cloud: Concept and Research challenges (IEEE, 2011), pp.42-46.
[26] Paul Marshall, Kate Keahey& Tim Freeman: Elastic Site (IEEE, 2010), pp.43-52.
[27] PenchengXiong,Yun Chi, Shenghuo Zhu, Hyun Jin Moon, CaltonPu&HakanHacigumus: Intelligent Management Of Virtualized Resources for Database Systems in Cloud Environment(IEEE,2011),pp.87-98.
[28] RerngvitYanggratoke, FetahiWuhib and Rolf Stadler: Gossip-based resource allocation for green computing in Large Clouds: 7th International conference on network and service management, Paris, France, 24-28 October, 2011.
[29] Richard T.B.Ma,Dah Ming Chiu and John C.S.Lui, Vishal Misra and Dan Rubenstein:On Resource Management for Cloud users :a Generalized Kelly Mechanism Approach.
[30] ShikhareshMajumdar: Resource Management on cloud: Handling uncertainties in Parameters and Policies (CSI communicatons, 2011, edn) pp.16-19.
[31] Shuo Liu Gang Quan ShangpingRen On –Line scheduling of real time services for cloud computing. In world congress on services, pages 459-464, 2010.
[32] T.Wood et al. Black Box and gray box strategies for virtual machine migration. In Proc 4th USENIX Symposium on Networked Systems Design and Implementation (NSDI 07), pages 229-242.
[33] Tram Truong Huu& John Montagnat: Virtual Resource Allocations distribution on a cloud infrastructure (IEEE, 2010), pp.612-617.
[34] WaheedIqbal, Matthew N.Dailey, Imran Ali and Paul Janecek& David Carrera: Adaptive Resource Allocation for Back-end Mashup Applications on a heterogeneous private cloud.
[35] Weisong Hu et al. : Multiple Job Optimization in Map Reduce for Heterogeneous Workloads: IEEE Sixth International Conference on Semantics, Knowledge and Grids 2010, pp.135-140.
[36] Wei-Tek Tsai Qihong Shao Xin Sun Elston, J. Service-oriented cloud computing. In world congress on services, pages 473-478, 2010.
[37] Wei-Yu Lin et al.: Dynamic Auction Mechanism for Cloud Resource Allocation: 2010 IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing, pp.591-592.
[38] X.Zhu et al. Integrated capacity and workload management for the next generation data center. In proc.5th international conference on Automatic computing (ICAC'08), pages 172-181, IEEE Press 2008.
[39] XiaoyiLu,Jian Lin, Li Zha and ZhiweiXu: Vega Ling Cloud: A Resource Single Leasing Point System to Support Heterogenous Application Modes on Shared Infrastructure(IEEE,2011),pp.99-106.
[40] Xiaoying Wang et al. : Design and Implementation Of Adaptive Resource Co-allocation Approaches for Cloud Service Environments : IEEE 3rd International Conference on Advanced Computer Theory and Engineering 2010,V2,pp,484-488.
[41] Xindong YOU, Xianghua XU, Jian Wan, Dongjin YU: RAS-M: Resource Allocation Strategy based on Market Mechanism in Cloud Computing (IEEE, 2009), pp.256-263.
[42] Y.C Lee et.al,"Project driven service request scheduling in clouds". In proceedings of the international symposium on cluster & Grid Computing. (CC Grid 2010), Melbourne, Australia.
[43] Yang wt.al A profile based approach to Just in time scalability for cloud applications,IEEE international conference on cloud computing, 2009,pp 9-16.

[44] Zhen Kong et.al: Mechanism Design for Stochastic Virtual Resource Allocation in Non-Cooperative Cloud Systems: 2011 IEEE 4th International Conference on Cloud Computing: pp, 614-621.
[45]ZhixiongChen,JongP.Yoon,"International Conference on P2P,Parallel, Grid, Cloud and Internet Computing",2010 IEEE:pp 250-257.