ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Rule Based Resource Allocation Model for Multi-Tier Applications

[I]Harjot Kaur Jhajj, [II]Ashish Jalota
[I]M.Tech Student, [II]Assistant Professor, CSE
[I,II]Desh Bhagat University, Punjab, India

Abstract
In a shared virtual computing environment, various capacity demands, quality of services, dynamic workload are required. Optimization of resources on the cloud computing is key to improve the efficiency of clod computing. Existing resource allocation methods mainly focus on either central global optimization or local optimization within a server, but with some limitations on the scalability of cloud. In the existing 2- tire on demand resource allocation mechanism consists of the global and local resource allocation with on- demand capacities to the concurrent applications. On- demand resource allocation is represented using optimization theory. SLA based resource allocation is proposed for the Multi-tier applications. Service bus and caching is used to store the previous handled request by the server. The caching improved the scalability of the applications. It also improves the load balancing and leveling of the applications by reducing the unnecessary allocation of resources even not required in the particular web page in a website.

Keywords
Cloud computing, Multi-tier Application, Resource Allocation, SLA

I. Introduction

Cloud computing is an on demand network model used for enabling convenient sharing of resources. It configures the storages, application, services, networks that can be used rapidly for management of service provider interaction. It connects the clients with the centralized data centre which is very far away[4]. We can access the data centre either by using the company network or via internet or sometimes both [2]. We can access applications on cloud by using mobile phones, a PC or through a tablet. There are many applications available on cloud centre and when user sends a request for a particular application it checks for that application in the cloud centre and if the application is available then the user gets access to it otherwise the request is declined. Cloud computing involves multiple components and applications communicating with each other through programming interfaces or application interfaces usually in web services and 3- tier architecture [3]. There are multiple programs that are run and executed together over a universal interface. The complexity in cloud computing is less and it is controlled. . Cloud computing may be stretched as much as we can but still it cannot handle all the traffic sent on the server by clients. Cloud computing increases scalability and performance. Through cloud computing web traffic can be fluctuated and the load of the server can be reduced. Users get fast results in very less time. Through clod computing the cost is greatly reduced and in a public cloud delivery model capital expenditure is converted to operational expenditure.

The presentation logic in three tier architecture is handled by the client. The communication between client tier and business logic is very little. Internet browser is example of thin client. The information provided by the browser is very fast and with no delay in processing. The application design today is complex software that is implemented on multitier architecture. Each tier provides a particular services to the next tier and user services from the previous tiers. To allocate resources for a multitier application is a complex task rather than assigning resources to single tier. The reason behind that is tiers are not homogenous. One tier can decrease the overall profit even if the other tiers have good Quality of Services.

The SLA which is established with the clients must fulfill the requirements given by the user. When a client demands for different application it becomes challenges to fulfill the need of the client at that particular periods of time so optimal resource provisioning is a bottleneck in SLA's [7]. Steps that can be taken to overcome these problems we have to reduce the cost incurred on the data centre operator and meet the requirements of the clients.

In the cloud computing there are three types of services. There are all services has a very different business value proposition.

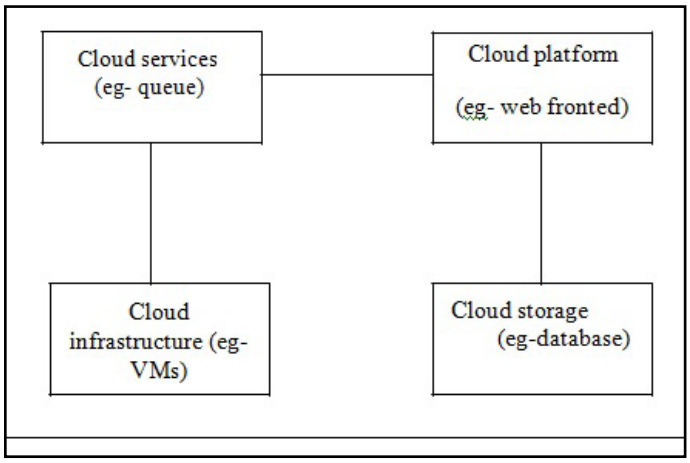


Fig. 1: General Architecture of cloud computing

II. Litereture Survey

Ving song, Yuzhong Sun, Member, et al 2013, in this paper two- tier architecture has been proposed. In which resources are allocated locally and globally with the feedback to provide on demand capacities to the concurrent applications. Based on the proposed dynamic resource allocation mechanisms, a set of on demand resource allocation algorithm has been proposed in this paper. This algorithm insures the performance of the critical application named by the data center manager when resource competition arises. According to the time varying capacity demands and the quality of application, in this paper they have shown 26% higher CPU utilization than traditional computing frame work, in which application used exclusive servers. the two-

tier on demand resource allocation further improve performance by 9 to 16 % for those critical applications 75% of the maximum performance improvement, introducing up to 5% performance degradation compares to others. They introduce a new model which replaced the previous model and tried to overcome the problems which were in the existing model. This model consists of two layers.

Avinash Mehta, et al (2011), Energy Conservation in Cloud Infrastructures International Institute of Information Technology, Bangalore [7]. This paper proposes the service request prediction model to achieving energy conservation in existing cloud infrastructure. The work of the service request prediction model is to determine the predefined period of time in which the server cluster will be underutilized. In this model they also define the load balancing mechanism. In this mechanism they accumulates all the requests, rather than distributing the load. This model also provides the less SLA violation with energy conservation. It reduces the overall cost and increases the lifetime of infrastructure.

Jianfeng Yan Wen-Syan Li SAP Technology Lab, Calibrating Resource Allocation for Parallel Processing of Analytic Tasks, 2009 [13], In this paper they described the challenge for the automated calibration of resource allocation for parallel processing and proposed an algorithm. This algorithm represented runtime statistic information and also calibrate the resource allocation accordingly. The experimental result of this algorithm describes that this algorithm is faster and more precision as compare to the other well know algorithms and also the pervious proposed algorithms.

Jinhua Hu, Jianhua Gu et al., A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment, 3rd International Symposium on Parallel Architectures, Algorithms and Programming, 2010 [4], In this paper they described how balance the load on the VMs resources. For this work they purpose genetic algorithm. This algorithm increases load balancing factor and reduce the dynamic migration and high migration cost. This algorithm performs better even load is stable variant [4]. In this paper they also used the mapping between the VMs and physical machines for load balancing. This paper builds a model based on the concrete situations of cloud computing. It considers the historical data and current states of VM, uses tree structure to do the coding in genetic algorithm, proposes the correspondent strategies of selection, hybridization and variation also puts some control on the method so that it has better astringency. However in real cloud computing environment, there might be dynamic change in VMs, and there also might be an increase of computing cost of virtualization software and some unpredicted load wastage with the increase of VM number started on every physical machine.

Karthik Kumar, et al. Resource Allocation for Real-Time Tasks using Cloud Computing, School of Electrical and Computer Engineering, Purdue University, West Lafayette, 2011 [5], According to this paper they purposed the method to allocate the resources for real- time tasks. They use the infrastructure as a service model. There is a condition; the real time task has to be completed in the particular time period and also before the deadline. For this problem they purpose a scheme that is EDF- greedy scheme. According to this scheme they consider the temporal overlapping to allocate resources efficiently.

Chongmin Li, Dongsheng Wang, Haixia Wang, Yibo Xue, This paper tells the concept, which says that memory scheduling algorithms should be designed to handle the memory requests from different threads. This can provide better system throughput and the fairness in the working of the system. A new algorithm known as "priority based fair scheduling is discussed in which it is said that in it classifies threads memory access behavior by dynamically updated priorities. Here it says that the threads those are latency sensitive they have top priority for giving the throughput to the system.

Kazuki MOCHIZUKI and Shin-ichi KURIBAYASHI, Evaluation of optimal resource allocation method for cloud computing environments with limited electric power capacity, 2011 International Conference on Network-Based Information Systems [8], In this paper they says that the limitation on the "electric power capacity" is major concept in each area, so they focus on that how they allocate the resources to the cloud computing with the limited electric power capacity. They say that:

a. Network bandwidth and processing ability both are allocated simultaneously.
b. They also purpose a method for optimally allocating the bandwidth and processing ability as well as the electric power capacity.
c. They also purpose an algorithm for the electric power consumption. This algorithm reduces the electric power consumption by aggregating requests of multiple areas.

Tino Schlegel, Ryszard Kowalczyk, Quoc Bao, Decentralized Co-Allocation of Interrelated Resources in Dynamic Environments, 2008, [11], In this paper they mentioned the decentralized co-allocation of interrelated resources in dynamic environment and it also includes the repeated jobs in real time. There is a resource broker agent that is autonomously allocating the resources for the execution of jobs but allocates the resources by the resource broker agent. It is based on the individual feedback and that feedback is received from the previous resource allocation decision. The result of this algorithm is very good and efficient for the open and dynamic environment with real application. There is a factor deadlock that may occur in between the agents, so for this factor they also purpose randomising techniques. They say that a limitation is set on the number of suitable resources providers in each broker.

T.R. Gopalakrishnan Nair, Vaidehi M, Efficient resource arbitration and allocation strategies in cloud computing through**,** 2011 [9]**,** in this paper they purposed an is rule based resource allocation (RBRA). This algorithm is based on the queuing model, means it is based on the priority management and also the FIFO approach that is first in first out approach. It can be said that, there is optimal resource allocation which is occurring if the rate of resource request from all subscribers is less than the rate with which the resource is allocated to subscribers.

Sai Prashanth Muralidhara, Lavanya Subramajan, Onur Mutlu, Mehmut Kandemir and Thomas Moscibroda. The second thing in research is to know how to partition the single memory channel. So here it is discussed about what is channel partitioning. So, first of all it is said that the performance benefits of mapping the pages of applications with largely different memory intensities to separate channels.

Conventional page mapping - in conventional page mapping the requests have to wait until the earlier requests are processed. Though the requests coming from the second source are creating a disturbance to the other but still it continues to run the processes which have arrived earlier.

Channel Partitioning - in this approach the requests that are coming do not wait and are processed at the time they arrive. So there

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

latency of all the requests are eliminated and thus the there is increase in the processing.

MCP consists of three steps which are performed at different intervals:

Profiling of an application - under this the statistics related to MPKI and RBH are collected and it is seen that due to which application how much harm is caused to the other applications.

Assignment to preferred channel - then after the profiling the applications are assigned preferred channels, the goal of this is to (a) separate low memory- intensity applications from that of a high memory-intensity applications, (b) among the high memory-intensity application, data of low row-buffer locality applications from that of high row-buffer locality applications.

Vincent C. Emeakaroha, Ivona Brandic, Michael Maurer, Ivan Breskovic, SLA-Aware Application Deployment and Resource Allocation in Clouds, 2011[1], There is a one parameter known as SLA that is considered. In this paper they describe the multiple SLA parameters for deploying application in clouds. They define the heuristic design and implementation also. The heuristic design includes the load balancing mechanism. They also include the flexible on- demand resources usage in the heuristic. The aim of heuristic scheduling is to schedule applications on VMs with SLA terms and deployment. VMs on physical resources are totally based on resource availability.

Hao Li, Jianhui Liu,Guo Tang, A Pricing Algorithm for Cloud Computing Resources**,** 2011 International Conference on Network Computing and Information Security [9]. This paper focuses on the scheduling and optimization of physical resources but they can't provide the physical resources without the economic principles in the cloud applications. They purposed a cloud banking model. They consider the operating mechanism in banks, classification and quantification for the cloud resources, quality of services, and quality of use of cloud resources parameter. They also defined the pricing algorithm. The core of algorithm is CRP, CRP describe the following services

It obtains the described tasks from the agent and participates in the competition.

Chrysa Papagianni, Aris Leivadeas, Symeon Papavassiliou, Vasilis Maglaris, Cristina Cervello´ -Pastor, and _ Alvaro Monje, On the Optimal Allocation of Virtual Resources in Cloud Computing Networks, 2013. [3], Cloud computing is by building advances on virtualization and distributed computing to support cost-efficient usage of computing resources, emphasizing on resource scalability and on demand services. In this paper they are providing the unified resources allocation framework for networked clouds. They firstly formulated the optimal networked cloud mapping problem as a MIP (mixed integer programming problem). Efficiently mapping of resource requests onto a shared substrate interconnecting various islands of computing resources and adopt a heuristic methodology [3]. IaaS provides the on demand and immediate resources, actually the computing resources with the cost saving according to the user. Cloud provides two keys as Cloud computing and Networking. Functional parameter defined characteristic and properties of computing/ networking resources, for example operating system, supporting virtualization environment. Non- functional parameter specifies the criteria and the constraint, for example maximize the number of interfaces for each node, maximum disk space at the end.

Yan, Jianfeng, and Wen-Syan Li. "Calibrating Resource Allocation for Parallel Processing of Analytic Tasks." In e-Business Engineering, 2009. In this paper they defined the traditional model for computing. As this model have certain advantages and certain disadvantages also. In this model there was no on-demand allocation of resources and the workload factor was ignored which results in the slow processing. This model works on dedicated resources which means that only limited resources where allocated to it and due to these factors the processing rate degrades. Secondly there came a new model which replaced the previous model and tried to overcome the problems which were in the existing model. This model consists of two layers.

III. Problem Formulation

Resource allocation mechanism should also consider the current status of each resource in the cloud environment. When apply any algorithm for better allocation of physical and/ or virtual resources the aim is to minimize the operational cost of the cloud environment. The first problem in the resource allocation is if the request is coming then how the resources are modelled. A networked cloud environment and request mapping model was designed which we also call as hardware representation. In this model the requests are coming from the user and going to the applications. Node mapping and link mapping is used in this architecture is to allocate the resources virtually over the cloud.

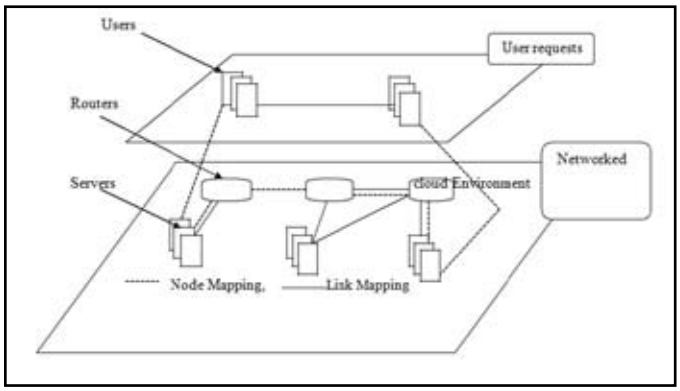
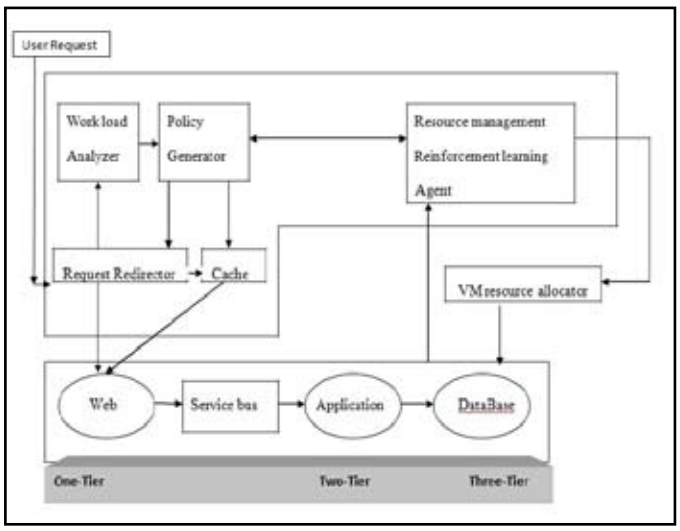


Fig. 2: Hardware representation of on-demand resource allocation problem

The problem is how the cloud IaaS handles the request, efficiently, efficient mapping of user requests for virtual resources onto a shared substrate interconnecting isolated island of computing resources with multitier application, SLA, caching and service bus. The problem is to solve the real- time problem of mapping virtual resources to substrate resources with limited assets. The virtual nodes and virtual links are known as virtual network embedding, but the main problem is, how the overload on each tier is distributed efficiently.

IV. Proposed Model



Working of Proposed Architecture

Step 1: Request Redirector
First user request will come to the request redirector.
Step 2: Work Load Analyzer
It will recognize the website:

1. One tier
2. Two tier
3. Three tier

The working of work load analyzer is to identify the type of request coming to the server.

Step 3: Policy Generator

Policy generator will identify the policy according to the website.

1. Different policies to handle (1,2,3 Tier applications)
2. Access permission of users
3. Security
4. Internet Bandwidth
5. Type of user

Step 4: Resource Management Reinforcement Learning Agent
The working of this agent is to learn the different resource allocation policies according to the type of request. It will store the previous resource allocation for the type of request coming from user and type of request it will handle on usually basis.
Step 5: VM Resource Allocator
It will allocate the resources and required VM according to the application. Find the cheapest way to handle the request.

1. One Tier can be of (window or linux)
2. Two Tier depends (Microsoft, LAMP, Phython)
3. Three tier depends (Which database)

Step 6: Service Bus
It will be added to the Client-Server communication at server side which will store some previous accessed data and directly connected with the cache. It will enhance the speed at cloud.

V. Conclusion

In this paper, one tier application only requires the HTML, CSS and JavaScript services, which can be easily provided by the low cost servers such as Linux or open source operating system. Two-tier application or web page requires web server and application oriented VM. Again it can further optimize between windows and Linux. Three tier application or web page requires all services but it further has scope of optimization. So not taking Multi-tier application as a whole working on the page inside the website, how many 1-tier, 2-tier and 3-tier. Further learning agent are used to learn the behaviour of different websites by time for further optimization and Policy generator is decide to allocate the resources for individual page request. Overall the Multi-tier application performance has been improved.

References

[1]. *Emeakaroha, Vincent C., Ivona Brandic, Michael Maurer, and Ivan Breskovic. "SLA-Aware application deployment and resource allocation in clouds." In Computer Software and Applications Conference Workshops (COMPSACW), 2011 IEEE 35th Annual, pp. 298-303. IEEE, 2011.*

[2]. *Goudarzi, Hadi, and Massoud Pedram. "Multi-dimensional sla-based resource allocation for multi-tier cloud computing systems." In Cloud Computing (CLOUD), 2011 IEEE International Conference on, pp. 324-331. IEEE, 2011.*

[3]. *Guo, Yanfei, Palden Lama, Jia Rao, and Xiaobo Zhou. "V-Cache: Towards Flexible Resource Provisioning for Multi-tier Appslications in IaaS Clouds."*

[4]. *Hu, Jinhua, Jianhua Gu, Guofei Sun, and Tianhai Zhao. "A scheduling strategy on load balancing of virtual machine resources in cloud computing environment." In Parallel Architectures, Algorithms and Programming (PAAP), 2010 Third International Symposium on, pp. 89-96. IEEE, 2010.*

[5]. *Kumar, Karthik, Jing Feng, Yamini Nimmagadda, and Yung-Hsiang Lu. "Resource allocation for real-time tasks using cloud computing." In Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on, pp. 1-7. IEEE, 2011.*

[6]. *Li, Hao, Jianhui Liu, and Guo Tang. "A pricing algorithm for cloud computing resources." In Network Computing and Information Security (NCIS), 2011 International Conference on, vol. 1, pp. 69-73. IEEE, 2011.*

[8]. *Mehta, Avinash, Mukesh Menaria, Sanket Dangi, and Shrisha Rao. "Energy conservation in cloud infrastructures." In Systems Conference (SysCon), 2011 IEEE International, pp. 456-460. IEEE, 2011.*

[9]. *Mochizuki, Kazuki, and Shin-ichi Kuribayashi. "Evaluation of optimal resource allocation method for cloud computing environments with limited electric power capacity." In Network-Based Information Systems (NBiS), 2011 14th International Conference on, pp. 1-5. IEEE, 2011.*

[10]. *Nair, TR Gopalakrishnan, and M. Vaidehi. "Efficient resource arbitration and allocation strategies in cloud computing through virtualization." In Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on, pp. 397-401. IEEE, 2011.*

[11]. *Papagianni, Chrysa, Aris Leivadeas, Symeon Papavassiliou, Vasilis Maglaris, and A. Monje. "On the optimal allocation of virtual resources in cloud computing networks." (2013): 1-1.*

[12]. *Schlegel, Tino, Ryszard Kowalczyk, and Quoc Bao Vo. "Decentralized co-allocation of interrelated resources in dynamic environments." In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, vol. 2, pp. 104-108. IEEE, 2008.*

[13]. *Song, Ying, Yuzhong Sun, and Weisong Shi. "A Two-Tiered On-Demand Resource Allocation Mechanism for VM-Based*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Data Centers." (2013): 1-1.

[14]. Yan, Jianfeng, and Wen-Syan Li. "Calibrating Resource Allocation for Parallel Processing of Analytic Tasks." In e-Business Engineering, 2009. ICEBE'09. IEEE International Conference on, pp. 327-332. IEEE, 2009.

Web References:

[15]. http://www.windowsazure.com/en-us/develop/net/tutorials/multi-tier-application/

[16]. http://www.academia.edu/3363069/Resource_Allocation_in_Clouds_Concepts_Tools_a nd_Research_Challenges