ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 1, January 2013

Priority based resource allocation model for cloud computing

K C Gouda, Radhika T V, Akshatha M

Abstract— **Cloud computing is a model which enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. A cloud environment consists of multiple customers requesting for resources in a dynamic environment with possible constraints. In the existing economy based models of cloud computing, allocating the resource efficiently is a challenging job. In this paper we propose a new approach that allocates resource with minimum wastage and provides maximum profit. The developed resource allocation algorithm is based on different parameters like time, cost, No of processor request etc. The developed priority algorithm is used for a better resource allocation of jobs in the cloud environment used for the simulation of different models or jobs in an efficient way. After the efficient resource allocation of various jobs, an evaluation is being carried out which illustrates the better performance of cloud computing with profit. A performance study of all the algorithms in various systems and case studies are also presented.**

***Index Terms*— Cloud Computing, Computing performance evaluation, Priority based algorithm, Resource allocation.**

I. INTRODUCTION

In the present advanced information and technology era, cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advent of this technology in the present age, the cost of computation, application hosting, content storage and delivery is reduced significantly. Cloud computing proved to be a practical approach to experience direct cost benefits and it has the potential to transform a data centre from a capital-intensive set up to a variable priced environment. Cloud computing has emerged as a popular solution to provide cheap and easy access to externalized IT resources. An increasing number of organizations (e.g., research centers, enterprises etc.) benefit from Cloud computing to host their applications. In contrast to previous paradigms (Clusters and Grid computing), Cloud computing is not application-oriented but service-oriented it offers on-demand virtualized resources as measurable and billable utilities. In other terms, Cloud computing provides access to IT resources as services ranging from direct access to hardware equipment to more sophisticated applications.
Cloud computing can be considered as an extension of grid computing. One of the main characteristics of cloud computing is on-demand self-service. That means Cloud computing characteristically has provision for on-demand IT resource allocation and instantaneous scalability. Unlike Grid computing that typically provides persistent and permanent use of all available IT resources, the cloud computing is very specific on the consumers demand, based on his current computing requirements and therefore eliminates over provisioning of available IT resources.

Primary advantage with the cloud computing is that the business enterprises can scale up to requisite capacities instantaneously without having to invest in new IT infrastructure that includes computers, network or database administrators and new licensed software.

The business enterprises can save huge amount of expense by avoiding build and manage large data centers for in house applications or data storage. The consumers of the cloud computing do not have to own the IT infrastructure and therefore need not care about maintenance of servers and networks in the cloud. They just pay for services on demand that is based on running of application instances normally varying depending upon use of Internet bandwidth, number of instances in action and amount of data transferred at specific time.

In this paper, we present the methods for efficient resource allocation that will help cloud owner to reduce wastage of resources and to achieve maximum profit. Efficient resource allocation in the cloud is a very challenging task as it needs to satisfy both the user's requirements and server's performance equally. Resource allocation in cloud computing environment is defined as assignment of available resources such as CPU, memory, storage, network bandwidth etc in an economic way. It is the main part of resource management. Yet, an important problem that must be addressed effectively in the cloud is how to manage Quality of Services (QoS) and maintain Service Level Agreement (SLA) for cloud users that share cloud resources.

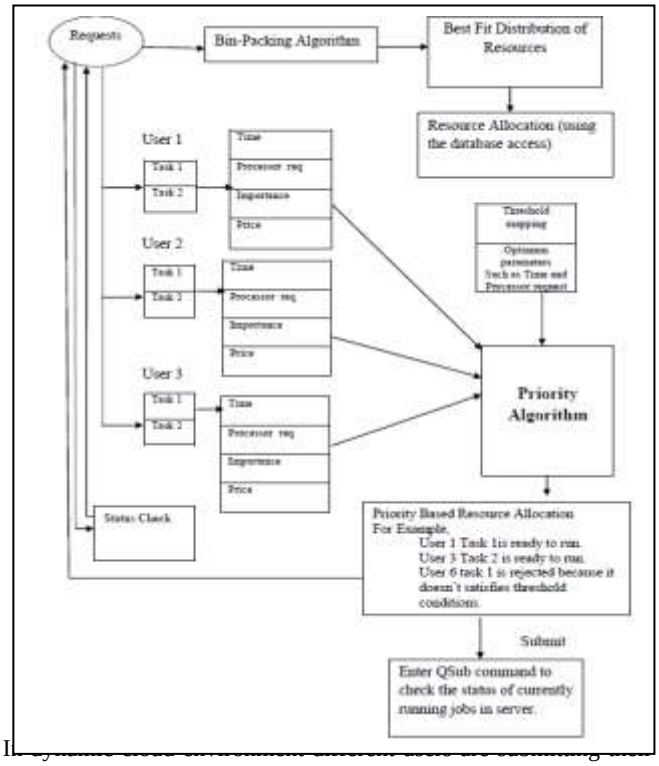
In this paper we have proposed priority algorithm that mainly decides priority among different user request based on many parameters like cost of resource, time needed to access, task type, number of processors needed to run the job or task etc. Finally Profit model algorithm is discussed that is basically used to calculate total profit gained by cloud owner by serving to all customer's request. It considers many parameters such as contract Length (ConLen), virtual machine cost (VMcost), price of each service (PriServ), service intiation time(iniTimeSev), penalty cost etc.

II. RESOURCE ALLOCATION MODEL

The Resource Allocation Model presented in this paper is basically an algorithm for efficient resource allocation in a cloud computing Environment. The Proposed model has been developed by considering various parameters such as cost, profit, user, time, Number of processor request, resource assigned, resource availability, resource selection criteria etc. In Resource allocation Model, clients are customers or users of cloud which sends service request that is client sends job request that is to be executed or run in cloud server. Server in cloud computing environment is the cloud service provider which will run the task or job submitted by client. The cloud administrator plays key role in efficient resource allocation because he decides the priority among the different user

request. This priority based resource allocation considers the parameters discussed above.

Virtualization is another important topic in cloud computing. It is a computing technology that enables a single user to access multiple physical devices. Another way to look at it is a single computer controlling multiple machines, or one operating system utilizing multiple computers to analyze a database. With cloud computing, the software programs that are used aren't run from your personal computer, but rather are Stored on servers housed elsewhere and accessed via the Internet. The resource allocation model that decides priority among different user request is shown in figure.



In dynamic cloud environment different users are submitting their request. Each request consists of different task. For each task different parameters are considered such as time, Processor request, Importance and price. Time refers to computation time needed to complete the particular task, Processor request refers to number of processors needed to run the task. More the number of processor, faster will be the completion of task. Importance refers to how important the user to a cloud administrator(admin) that is whether the user is old customer to cloud or new customer. Finally price parameter refers to cost charged by cloud admin to cloud users.

Earlier Bin-Packing algorithm were used for best fit distribution of resources in cloud environment. Bin Packing is a mathematical way to deal with efficiently fitting resources into Bins. A formal definition of the Bin Packing (BP) problem can be defined as given a list of objects and their weights, and a collection of bins of fixed size, find the smallest number of bins so that all of the objects are assigned to a bin. Now, a Bin is something that can hold inside itself a certain amount (it's Bin Height). Every Resource is of a certain, nonzero, and positive value (Resource Height).

Based on all the parameters considered above and also based on some threshold parameters, priority algorithm decides priority among different task submitted by different users. The user's task with higher priority will be given first chance to run. The user's task with next higher priority will be given second chance and so on. The task which exceed threshold will be aborted. Cloud admin can also check the status in order to know which the running tasks are and which are in queue. In this way by using priority algorithm, cloud administrator can efficiently allocate the resources among the users with minimum wastage and provides maximum profit.

III. PRIORITY ALGORITHM

In a cloud computing environment, multiple customers are submitting job request with possible constraints that is multiple users are requesting same resource. For example in a high performance computational environment which mainly deal with scientific simulations such as weather prediction, rainfall simulation, monsoon prediction and cyclone simulation etc which requires huge amount of computing resources such as processors, servers, storage etc. Many users are requesting these computational resources to run their model which is used for scientific predictions. So at this situation it will be problem for cloud administrator to decide how to allocate the available resources among the requested users.

Table 1. Parameters considered for job submission

No. of Users	eg: 10 users
Servers	eg: S1, S2,S3
Time to run	eg: 4 Hours
No Processors requested	eg: 8 Processors
Amount of memory	eg: 5 GB, 1 TB etc.
Time of request	eg:1:30 am
Software to be used	eg: Matlab, Grads,NetCDF
Job type	eg:Sequential or parallel.
User type	eg: Internal or External

The proposed priority algorithm helps cloud admin to decide priority among the users and allocate resources efficiently according to priority. This resource allocation technique is more efficient than grid and utility computing because in those systems there is no priority among the user request and cloud administrator is randomly taking decision and he is giving priority to those user who have submitted their job first that is based on first come first serve method. But with the advent of cloud computing and by using this implemented priority algorithm, the cloud admin can easily take decision based on different parameters discussed earlier to decide priority among different user request so that admin can efficiently allocate the available resources and with cost-effectiveness as well as satisfaction from users. The table 1 shows the parameters considered for job/task submission cloud computing environment.

Algorithm: *To compute and assign the priority for each request based on the threshold value and allocate the service to each request's.*
Step 1*: [Read the clients request data i.e, time, importance, price, node and requested server name]*
Insert all values into the linked list
Step 2: *[For each request and its tasks find the* ***time*** *priority value based on the predefined conditions]*
Assign priority value to each task for the client's request.
t_p[i] = priority valu
Step 3: *[For each request and its tasks find the* ***node*** *priority value based on the predefined conditions]*
Assign priority value to each task for the client's request.
n_p[i] = priority value;
Step 4: *[For each client's input data check whether it is within the threshold value or not]*
if (input value is within the threshold limit and total node <= available node)
[Add respective computed time and node priority value and other parameters like
importance and price]
Sum[k] = t_p[i] + n_p[i] + importance + price
Print —Ready to execute
available node = available node – total node
else if (input value is within the threshold limit)
sum[k] = t_p[i] + n_p[i] + importance + price
print —within the limit but it is in queue
else
print —Exeed the condition
Step 5: *[Sort the sum[k] values]*
Step 6: *Client's request is ready to execute from least values of sum[k]*
Stop

In order to run particular model huge computational resources such as server, memory in terms of storage disk, processors, software etc are needed. Also some jobs are to be executed in parallel and some others in sequential manner. In that situation job type is also very important parameter. In a cloud environment type of user that is whether the user is internal to a cloud (in case of private cloud) or he is external to cloud(in case of public cloud) is also another important parameter to be considered during job submission. So the developed priority algorithm discusses in detail how efficiently it will help cloud admin to decide or calculate priority among the user requests. After the successful execution of resource allocation algorithm, the jobs requested by users needs to be submitted.

The main difficulties in the resource allocation in a cloud system are to take proper decision for job scheduling, execution of job, managing the status of job etc. Apart from traditional best fit and bin packing algorithm in this paper an algorithm is developed for the job allocation in the cloud environment to be decided by the cloud administrator. Several parameters listed in the table 1 are considered for the priority based on the client and server requirements and requests by the users. In the present algorithm to decide the resource allocation in a better and impartial way, a technique based on threshold of all the parameters (both client and server side) is considered. For example the requested number of processors cannot be more than 20 etc. (server) and a job maximum run time will be 200 hrs (user).The step wise explanation of the above said algorithm is presented in Figure 2.

IV. PERFORMANCE EVALUATION

This section describes the scenarios considered for performance evaluation, performance metrics and experimental results. All experiments are performed on the various simulation carried out at high performance computing platform by different users. Mainly four systems or models have used for the various applications like monsoon simulations, cyclone simulation, atmospheric cloud simulation etc. These models are submittend or run in high performance computing (cloud) servers. To evaluate the computational time variability for a model (Meso scale model, MM5) using different virtual cluster instances, we show the time spent in computation for seven different runs measured using the integrated performance monitoring.

It can be seen that there is 30% variability seen in compute time which can be explained by the processor distribution required for each run as shown in Figure 3. In the first run the job is very slow because it has used only two processors on the other hand the seventh run spent least amount of time in the computation because there 18 processors were used. The most important point to be noted is that the communication pattern also performs based on the overall run time. The difference between maximum and minimum is 120 seconds.

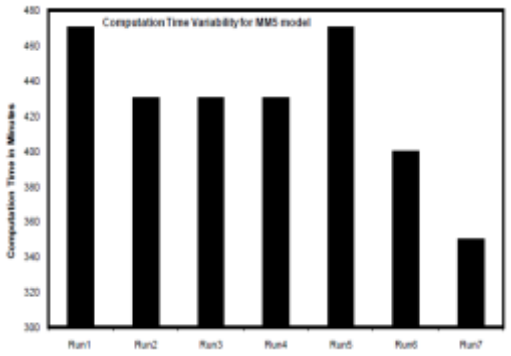


Fig 3. Computation Time Variability for MM5 model

The sustained performance per core for another application that is monsoon rainfall prediction using a General Circulation Model (GCM) is tested in different machines or servers (say A5, A4, A3, and A0).It is found that the performance is much better for the machine A5 in terms of sustained system performance as shown in Figure 4. So the main recommendation will be one should prefer to run in GCM in A5 rather than other machines.

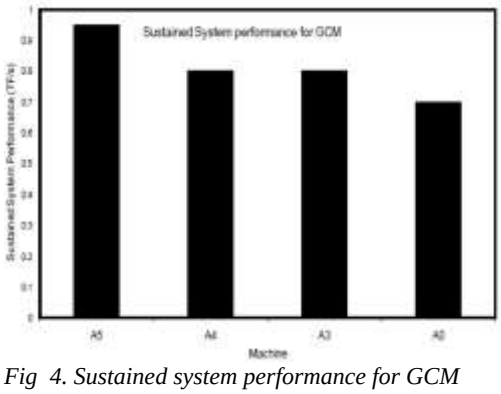


Fig 4. Sustained system performance for GCM

The same sustained performance is also carried out for different

applications namely process, MM5, Weather Research Forecast (WRF) model, Non Hydrostatic Model (NHM) and GCM in four separate servers A5, A4, A3, A0. It is found that the performance is different for each application in different machines. The performance of process model, WRF and MM5 machine A0 is high where as for NHM , A5 is performing well and for GCM both A5 and A0 are comparable as shown in the Figure 5.

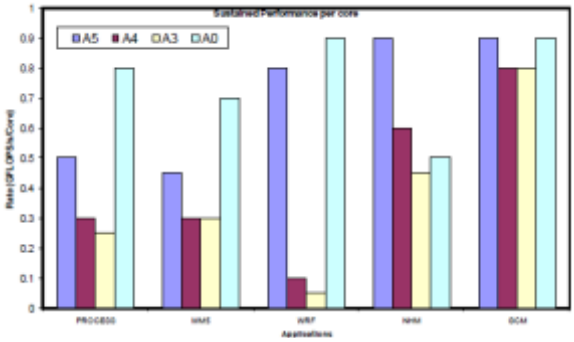


Fig 5. Sustained Performance per Core for different applications.

V. RELATED WORK

A large portion of the work in resource allocation in cloud computing mainly focused on the cost-effectiveness and easy maintenance of the systems [1]. Most of the work has been descriptive in nature. Patricia et al. [2] discusses this process in the context of distributed clouds, which are seen as systems where application developers can selectively lease geographically distributed resources. [2] Highlights and categorizes the main challenges inherent to the resource allocation process particular to distributed clouds, offering a stepwise view of this process that covers the initial modeling phase through to the optimization phase.

A critical evaluation of current network resource allocation strategies and their possible applicability in Cloud Computing Environment which is expected to gain a prominent profile in the Future Internet are presented in work by M. Asad Arfeen [3].

Atsuo Inomata et al. [4] has proposed a dynamic resource allocation method based on the load of VMs on IaaS, abbreviated as DAIaS. This method enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user.

It has been believed that a market-based resource allocation will be effective in a cloud computing environment where resources are virtualized and delivered to users as services (Fujiwara et al. [5]) and in such a market mechanism to allocate services to participants efficiently has proposed. The mechanism enables users to order a combination of services for workflows and coallocations and to reserve future/current services in a forward/spot market. The evaluation shows that the mechanism works well in probable setting.

In a cloud computing environment, it is necessary to simultaneously allocate both processing ability and network bandwidth needed to access it. Tomita et al [6] proposed the congestion control method for a cloud computing environment which reduces the size of required resource for congested resource type, instead of restricting all service requests as in the existing networks.

Mochizuki and Kuribayashi [7] presents cloud resource allocation guidelines in the case where there is a limit to electric power capacity available in each area, assuming a cloud computing environment in which both processing ability and network bandwidth are allocated simultaneously. Next, it proposes a method for optimally allocating processing ability and bandwidth as well as electric power capacity. Optimal allocation means that the number of requests that can be processed is maximized, and the power consumed by a request is minimized. It is demonstrated by simulation evaluations that the proposed method is effective.

VI. CONCLUSIONS

In this paper we have described a work on the allocation of resources in a dynamic cloud environment by using priority algorithm which decides the allocation sequence for different jobs requested among the different user after considering the priority based on some optimum threshold decided by the cloud owner. This resource allocation technique is more efficient than grid and utility. With the advent of cloud computing and by using this implemented priority algorithm the cloud admin can easily take decision based on the different parameters discussed earlier and can efficiently allocate the available resources and with cost effectiveness as well as satisfaction from users. Finally cloud admin will decide how much profit he can be gained by allocating the available resources and prioritizing among the different user request. Various case studies are presented in order to evaluate the performance of the algorithm in terms of sustained time for many applications in many servers of different configuration. Simulation results also shows that on average, the optimized resource allocation algorithm can be used for efficient cloud management.

VII. REFERENCES

[1] Website http://www.net-security.org/secworld.php?id=10886, Article on “Lack of admin rights mitigates most Microsoft vulnerabilities” Posted on 12 April 2011.

[2] Patricia Takako Endo, Andre Vitor de Almeida Palhares, Nadilma Nunes Pereira, 2011. Resource Allocation for Distributed Cloud: Concepts and Research Challenges, IEEE, july 2011.

[2] Hadi Goudarzi and Massoud Pedram University of Southern California, MaximizingProfit in Cloud Computing System via Resource Allocation.

[3] M.Asad Arfeen, Krzysztof Pawlikowski, Andreas Willig .2011, A Framework for Resource Allocation Strategies in Cloud Computing Environment, 2011 35th IEEE Annual Computer Software and Applications Conference Workshops.

[4] Atsuo Inomata, Taiki Morikawa, Minoru Ikebe. 2011, Proposal and Evaluation of a Dynamic Resource Allocation Method based on the Load of VMs on IaaS, IEEE 2011.

[5] Ikki Fujiwara, Isao ono, Kento Aida, Applying Double-sided Combinational Auctionsto Resource Allocation in Cloud Computing, 2010 10th Annual International Symposium on Applications and the Internet.

[6] Takuro TOMITA and Shin-ichi KURIBAYASHI, Congestion control method with fair resource allocation for cloud computing Environment . IEEE 2011.

[7] 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.

K C Gouda is currently working as a Scientist at CSIR Centre for Mathematical Modeling and Computer Simulation (CSIR C-MMACS). His research and professional career spans about twelve years of research and capacity building in modeling and computer simulation, satellite data processing, numerical modeling, Data mining, Data assimilation, cloud computing knowledge engineering and related subjects. His expertise is primarily in the domains of Software development for modeling and simulation. He is presently involved in several international and national projects related to HPC enabled modeling for weather and climate forecasting and analysis. He has published about 50 peer-reviewed papers as journal articles, book chapters, Technical reports and contributions to conference proceedings. He is a member of IMS, IEEE, AGU, AOGS and EGU. He is also a member in the board of studies of Department of Computer Science in the Jain University, Bangalore. He obtained his M.Sc., M.Phil, from Berhampur University, MCA from IGNOU, New Delhi and completed PhD from Mangalore University. He has Guided six M.Tech., 50 Masters and 20 B.E students for their academic project.

Radhika T V is currently working as a Lecturer in Dayananda Sagar College of Engineering, Bangalore. Her research interest includes cloud computing, Distributed Database, Software engineering and Data Mining. She obtained B.E (Computer Sc.) during 2010 and M.Tech degree in Computer Science & Engineering in 2012 from Visvesvaraya Technological University (VTU), Karnataka. She did her M.Tech thesis work at the CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore during 2011-2012. She has published some technical papers in conference proceedings.

Akshatha M is currently working as a Lecturer in Coorg Institute of Technology, Coorg, Karnataka. Her research interest includes cloud computing, Distributed Database, Software engineering and Data Mining. She obtained B.E (Computer Sc.) during 2010 and M.Tech degree in Computer Science & Engineering in 2012 from Visvesvaraya Technological University (VTU), Karnataka. She did her M.Tech thesis work at the CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore during 2011-2012. She has published some technical papers in conference proceedings.