QoS Supported SLA for Profit Maximization of Multiserver Configuration in Cloud Computing

Dr. P. Balakumar
Department of Computer Science
Mahendra Institute of Technology,Mallasamudaram,
Tiruchengode ,Tamil nadu, India

Deepa. V
Department of Computer Science
Mahendra Institute of Technology,Mallasamudaram,
Tiruchengode ,Tamil nadu, India

Abstract: **Cloud is one of the emerging technologies in computer engineering. Several companies move around on the way to this technology due to lessening maintenance cost.Numerous organizations offer cloud service such as SaaS, IaaS, PaaS. Different organization provides same service with different service charges and waiting time. So customers can select services from these cloud providers according to their criteria like cost and waiting time. By using 'demand pricing' strategy, providers can provide services with minimum cost without loosing any income or valuable resource time. But the existing system does not provide any automated job scheduling considering consumer cost, provider benefit, consumer waiting and provider idle time. This paper proposes a multi objective genetic algorithm for solving this multivariable optimization problem. This system provides a new cloud brokering mechanism with cloud service discovery using this optimization technique. This paper considers IaaS. In this system user submit a job to cloud. Cloud provides infrastructure to run this job and gave output to user. Here aim of user is to obtain output with minimum time and minimum cost. At the same time aim of provider is to increase the income. For that provide run more job within unit time. So We have to minimize consumer cost, consumer waiting time and provider idle time, and maximize provider benefit.**

Keywords— Cloud computing, multiserver system, pricing model, profit, response time, ,service charge, SLA, waiting time, server configuration, QoS

I. INTRODUCTION

The cloud is a next generation platform that provides dynamic pools of resource , virtualization, and high accessibility. Today, we utilize scalable, distributed computing environments within the margins of the Internet put into practice known as cloud computing. Cloud computing helps to decipher the daily computing problems, of hardware resource and software availability unhurried by computer users. The cloud computing provides an straightforward and non ineffectual solution for daily computing. Prevailing cloud systems mainly focus on finding an effective solution for the resource management.

Cloud Computing is Internet based computing where virtual shared servers provide software, infrastructure , devices, platform and other resources .They also provide hosting to customers on a pay-as-you-use basis. The cloud makes it possible for user to access your information from anywhere at any time. Cloud Computing enables a user what you need and pay for what you use model. This will enable businesses to invest on innovative solutions that will help them address key customer challenges instead of worrying about operational details.

"Cloud computing is a model that enable suitable, on-demand network access to a pool of shared configurable computing resources (e.g., servers, storage, networks, applications, and services) that able to be rapidly provisioned and released with minimal management effort or service provider interaction."

More specifically, cloud describes the use of a set of information, services, applications, and infrastructure . Infrastucture in cloud mainly consist of a great pool of storage ,computer, network, information, and resources. These can be rapidly scaled up or down providing an on-demand utility-like model of allocation and consumption. Cloud enhances collaboration, scaling, availability, and provides the potential for cost reduction through optimized and efficient computing.

Dynamic resource management is one of the crucial problems in the existing cloud environment due to changing resource demands of large computational tasks. They require knowledge of the resource needs for service requests of various kinds, and these needs may change over time. The blend of heterogeneous computing and cloud computing is emerging as a influential new standard to meet the requirements for high-performance computing (HPC).Cloud-based, heterogeneous[8] computing represents a significant step toward solving large computational tasks. Efficient load balancing[21] in a cloud is challenging since running machines have the problem of load imbalance due to resource variation in heterogeneous environment. For heterogeneous systems nodes have different processing capabilities, dynamic load balancing methods are preferred. This approach makes load balancing decision based on the current load status which varies on each machine . distribute load on the nodes at run time Modern parallel computing hardware demands increasingly specialized attention to the details of scheduling and load balancing across heterogeneous execution resources in cloud.

This research work mainly concentrating on developing an efficient dynamic load balancing method in heterogeneous cloud environment and hence achieve effective utilization of

resources and thereby increase the performance of the system. , we propose a new framework for calculation provider

II PROPOSED SYSTEM

The pricing model of a service provider in cloud computing is based on two components, namely, the income and the cost. For a service provider, the income is the service charge to users, and the cost is the renting cost plus the utility cost paid to infrastructure vendors. A pricing model in cloud computing includes many considerations, such as the quantity of a service (the amount of a service), the application environment's workload, of a multiserver system's configuration(the size and the speed), the penalty cost of a low-quality service, the consumer's satisfaction (the expected service time),the SLA, the quality of a service (the task waiting time and the task response time), , the cost of renting and energy consumption, the service provider's margin and profit. The profit (i.e., the net business gain) is the income minus the cost. In order to maximize the profit, it is essential that a service provider should be aware of both service charges and business costs, and also, how they are resolute by the of the application's characteristics and the multiserver system's configuration..

A.*The Model*

This system provide a pricing model, that takes following factors such as the workload of each and every application environment, the multiserver system's configuration, , the SLA , the approval of a consumer, the penalty cost of a low-quality service , the amount of a service, the QoS parameters of a service, the cost of renting, the cost of energy consumption.It also considers the service provider's margin and profit too.The Multiserver system is treated as an M/M/m queuing model, so as to solve the profit maximization problem analytically.

Main objective of this project is to allocate user submitted job to one of these N servers based on following criteria

1. Power consumption
2. Quality of service
3. Consumer satisfaction and mainly maximum profit

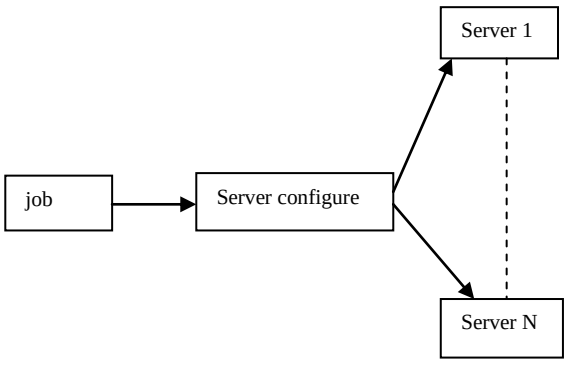


Figure 1:Architecture

benefit ,consumer benefit, consumer cost , quality of service and power consumption.

To achieve the objective of adaptive resource allocation for satisfying the service requests of customers, we use the architecture namely cloud booster archietecture.

B. M/M/m queuing model

The multiserver system is treated as an M/M/m queuing model, so as to formulate and solve the optimization problem can analytically . The system considers two server speed and power consumption models. They are the idle-speed model and the constant-speed model. The the waiting time of a newly arrived service request is obtained by calculating the probability density function . from that function the expected service charge to a service request is calculated. From the expected service charge to a service request calculate the expected net business gain in single unit of time, and thus determine the optimal server size and the server speed numerically.

Ant Colony Optimization

This section describes the ACO algorithm, which can be used for proper scheduling . Ant colony optimization is based on the technique known as Swarm Intelligence, which is a part of Artificial Intelligence.The ACO system contains two rules: 1. Local pheromone update rule, which applied whileconstructing solutions. 2. Global pheromone updating rule, which applied after all ants construct a solution.Furthermore, an ACO algorithm includes two more mechanisms: trail evaporation and, optionally, daemon actions. Trail evaporation decreases all trail values over time, in order to avoid unlimited accumulation of trails over some component. Daemon actions can be used to implement centralized actions which cannot be performed by single ants, such as the invocation of a local optimization procedure, or the update of global information to be used to decide whether to bias the search process from a non-local perspective At each step, each ant computes a set of feasible expansions to its current state, and moves to one of these in probability. The probability distribution is specified as follows. For ant k, the probability of moving from state t to state n depends on combined checking of two values :• the ant movement attractiveness , it is computed by some heuristic indicating the priori desirability of that move;• the movement , this indicates how capable it has been in the past to make that particular move: therefore it represents a posteriori indication of the movement desirability.

D. *Cloud Booster Architecture*

We use the following architecture To achieve the objective of adaptive resource allocation for satisfying the service requests of customers **.**It consist of mainly the following

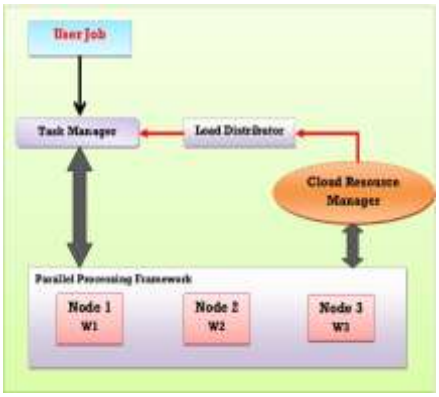


Figure 2: Cloud Booster Architecture

Users/Brokers: Users or brokers acting on their behalf submit service requests to the cloud via cloud controller for processing.

Cloud Controller: It acts as the interface between the cloud service provider and external users/brokers. It acts similar to the Queen in the ant colony. Virtual Machines(VMs): This is where the applications of customers will be deployed. We can dynamically create, start, stop and migrate these VMs depending on our requirement, from one physical machine to another. Physical Machines: These are the physical computing servers that will provide hardware infrastructure for creating virtual machines**.**

Cloud controller maintains a queue(Q) for storing the service requests for hosting the applications. It enqueues each of the service request received, in this queue. It generates the tester, scout, cleaner and worker ants periodically. The movement of these ant agents is modelled in the following way.

Each ant except Queen & Worker maintains a Visited Node list which is initially empty. Each node in the cloud maintains a list of neighbouring node's information. Whenever an ant reaches a node, it updates the controller about the current utilisation and randomly chooses an unvisited neighbouring node. When all the nodes are covered, it makes the Visited Node list empty and continues again in the same way.

We can change the number of ants that will be produced so that it will yield better results depending on our requirement. The next subsection describes the method used by worker ants for accepting or rejecting the service requests.

Worker ant:

Whenever a service request received in the queue, one of the worker ants creates a VM with a specific CPU processing power and memory etc, if accepted. So, worker ants are always looking in the queue to check if there are some pending requests to be processed. If such a request is found, it dequeues the request and calls Algorithm 1.

Since most of the CPUs are work conserving, we are creating a VM with specific CPU processing power and memory. Depending on the load, more intensive applications can use the resources of the other VMs having less load. The worker ant is only responsible for deploying the request on a VM . Load balancing decisions are taken by tester ant. After deploying, it creates a Service Level Agreement(SLA) monitor agent that monitors the hosted application. In the next subsection, we provide the details about the SLA monitor agent.

SLA Monitor Agent:

It calculates the Avg. response time and throughput of the hosted application by continuously monitoring it. It passes this information to the hypervisor on that host in the form of a variable(SLAM) which is calculated depending on the performance of the application

III. DISCUSSION

Like all business, the pricing model of a service provider in cloud computing is based on two components, namely, the income and the cost. For a service provider, the income is the service charge to users, and the cost is the renting cost plus the utility cost paid to infrastructure vendors. A pricing model in cloud computing includes many considerations, such as the amount of a service (the requirement of a service), the workload of an application environment, the configuration (the size and the speed) of a multiserver system, the service-level agreement, the satisfaction of a consumer (the expected service time), the quality of a service (the task waiting time and the task response time),the penalty of a low-quality service, the cost of renting, the cost of energy consumption, and a service provider's margin and profit. The profit (i.e., the net business gain) is the income minus the cost. To maximize the profit, a service provider be supposed to identify with both service charges and business costs, and in exacting, how they are resolute by the characteristics of the applications and the configura-tion of a multiserver system.

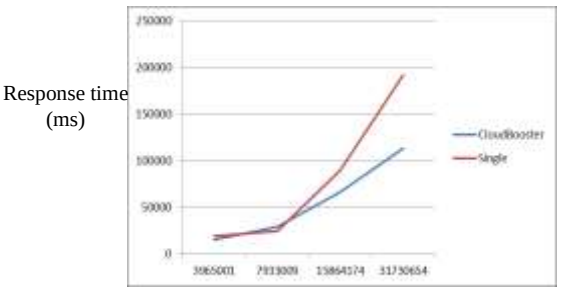
. Major Cloud computing companies have started to integrate frameworks for parallel data and making it easy for customers to access these. However, the processing frameworks which are currently used have been designed for static, homogeneous cluster setups

and disregard the particular nature of a cloud. Consequently, the allocated compute resources may be inadequate for big parts of the submitted job and unnecessarily increase processing time and cost. One of an IaaS cloud's key features is the provisioning of compute resources on demand. New VMs can be allocated at any time through a well-defined interface and become available in a matter of seconds

Existing parallel processing framework like Hadoop or Nephele did not have a better load balancing mechanism and prevailing load balancing approaches does not reflects the heterogeneity aspects of parallel machines.

Swarm intelligence (SI) is based on self-organized system's collective behavior .Stochastic Diffusion Search (SDS), Bacteria Foraging (BF), Ant Colony System (ACS), the Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), etc comes in swam intelligence.Most efficient one that we fount to implement for scheduling is ACO,which is implemented using CLOUD BOOSTER .

This research work mainly concentrating on developing an efficient dynamic load balancing method in heterogeneous cloud environment and hence achieve effective utilization of resources and thereby increase the performance of the system. Heterogeneous parallel processing framework is used as the deployment scenario and the proposed load balancing algorithm is incorporated into this system. The entire framework is named as "Cloud Booster". . We can expect efficient and more accurate result from the proposed scheme. Based on the capabilities of the system which is explained above, an expected performance graph of the system can be obtained as follows :



IV.CONCLUSION

Most of the firms are moving to cloud environment now a days. Moving to cloud is clearly a better alternative as they can add resources based on the traffic according to a pay-per-use model. The challenge faced in cloud computing is dynamically allocating resources to the users based on their demands. Resource management is critical component in this paradigm. Fundamental to these issues is the issue of load balancing. Inorder to to achieve a high user satisfaction and resource utilization ratio, it is essential to distribute the excess dynamic local workload evenly to all the nodes in the whole Cloud

Cloud booster system achieved better performance in the utilization of resources and response time of task is considerably reduced. This work proposes an effective solution for dynamic load balancing in dynamic and heterogeneous cloud environments. Deployment of this method works effectively in heterogeneous parallel processing system. It can be also used in large scale cloud environments that hosted sites.

Improvements can still be made in this work. Considering other resource such as bandwidth for node weight calculation is one of the future enhancements. Including threshold for job makes performance far better. In the future, load distribution process can be further modified for better allocation when there is an increase in the load and no of nodes in the system.

REFERENCES

[1] Junwei Cao, Kai Hwang, Keqin Li, and Albert Y. Zomaya, (JUNE 2013)" Optimal Multiserver Configuration for Profit Maximization in Cloud Computing" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 6
[2] http://doi.ieeecomputersociety.org/10.1109/TPDS.2012.203.
[3] Hadi Goudarzi and Massoud Pedram(2013)"Profit-Maximizing Resource Allocation for Multi-tier Cloud Computing Systems under Service Level Agreements" Large Scale Network-Centric Computing Systems, Wiley Series on Parallel and Distributed Computing, Jul.
[4] Hadi Goudarzi and Massoud Pedram(2009)"Maximizing Profit in Cloud Computing System via Resource Allocation"
[5] Seunghwan Yoo, and Sungchun Kim(2013)"SLA-Aware Adaptive Provisioning Method for Hybrid Workload Application on Cloud Computing Platform" Proceedings of the MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong
[6] D.E. Irwin, L.E. Grit, and J.S. Chase,(2004) "Balancing Risk and Rewarding a Market-Based Task Service," Proc. 13th IEEE Int'l Symp. HighPerformance Distributed Computing, pp. 160-169
[7] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger,(2007) "Economic Models for Resource Management and Scheduling in Grid Computing," Concurrency and Computation: Practice and Experience, vol. 14, pp. 1507-1542.
[8] Junwei Cao, Keqin Li, Ivan Stojmenovic, (2013)" Optimal Power Allocation and LoadDistribution for Multiple Heterogeneous Multicore Server Processors across Cloudsand Data Centers" IEEE TRANSACTIONS ON COMPUTERS
[9] P. Mell and T. Grance,(2009) "The NIST Definition of Cloud Computing,"Nat'l Inst. of Standards and Technology, http://csrc.nist. gov/groups/SNS/cloud-computing.
[10] C.S. Yeo and R. Buyya,(2006) "A Taxonomy of Market-Based Resource Management Systems for Utility-Driven Cluster Computing,"Software - Practice and Experience, vol. 36, pp. 1381-1419.
[11] Nada M. A. Al Salami(2009)"Ant Colony Optimization Algorithm" UbiCC Journal, Volume 4, Number 3, August
[12] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen,(1992) "Lowpower CMOS digital design," IEEE Journal on Solid-State Circuits, vol. 27, no. 4, pp. 473-484.
[13] Jon oberheide,EvanCooke,,FarnamJahanian(2008)"Empirical Exploitation of Live Virtual Machine Migration"
[14] Gong Chen, Wenbo He, Jie Liu, Suman Nath, Leonidas Rigas, Lin Xiao, Feng Zhao(2008) **"**Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services"USENIX association NSDI '08: 5th USENIX Symposium on Networked Systems Design and Implementation
[15] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, Andrew Warfield(2008)"Live Migration of Virtual Machines" USENIX Association NSDI '05: 2nd Symposium on Networked Systems Design & Implementation
[16] Utkarsh Jaiswal, Shweta Aggarwal(2011)"Ant Colony Optimization" International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011 1 ISSN 2229-5518
[17] Daniel Warneke, D and O. Kao(2011)."Exploiting Dynamic resource allocation for efficient parallel data processing in the cloud. IEEE Trans on Parallel Distributed Systems.
[18] M. Armbrust et al(2009)., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, Feb.
[19] N.Ani Brown Mary(2013)"Profit Maximization for Service Providers using Hybrid Pricing in Cloud Computing" International Journal of Computer Applications Technology and Research Volume 2– Issue 3, 218 – 223.
[20] Bryan Clark, Todd Deshane, Eli Dow, Stephen Evanchik, Matthew Finlayson, Jason Herne, Jeanna Neefe Matthews(2004)" Xen and the Art of Repeated Research"

[21] Ratan Mishra and Anant Jaiswal(2012)" Ant colony Optimization: A Solution of Load balancing in Cloud" International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April

[22] Ricardo Lent(2011)" Evaluating the Performance and Power Consumption of Systems with Virtual Machines" Third IEEE International Conference on Cloud Computing Technology and Science

[23] Vincent C. Emeakaroha, Ivona Brandic, Michael Maurer, Ivan Breskovic(2011)" SLA-Aware Application Deployment and Resource Allocation in Clouds"

[24] Junliang Chen, Chen Wang, Bing Bing Zhou, Lei Sun, Young Choon Lee, Albert Y. Zomaya(2011)" Tradeoffs between Profit and Customer Satisfaction for
Service Provisioning in the Cloud"

[25] Xiaorui Wang, Member, IEEE, and Yefu Wang, Student Member, IEEE(2011)" Coordinating Power Control and Performance Management for Virtualized Server Clusters" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 22, NO. 2, FEBRUARY

[26] H. Khazaei, J. Misic, and V.B. Misic,(may 2012) "Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 5,pp. 936-943.

[27] Dr.P.Balakumar,Department of computer science, Mahendra Institute of Technology,Mahendrapuri Namakkal,Tiruchengode,Tamilnadu.

[28] Deepa.V,Mahendra Institute of Technology,Mahendra Institute of Technology,Mahendrapuri,Namakkal Tiruchengode,Tamilnadu.