Dynamic Resource Allocation based on Priority in Service Oriented Systems

D.Sivapriyanka,
M.E-CSE,
CIET,

S.Santhanalakshmi,
ASST PROF/DEPT OF CSE,
CIET,

S.Gunasekaran, Ph.D
PROFESSOR& HOD,
DEPT OF CSE,
CIET,

ABSTRACT

Cloud Computing is a newly evolving platform that can be accessed by users as a service. It is used as storage for files, applications and infrastructure through the Internet. User can access everything as a service in on-demand basis named as pay-as-you-go model. Service-oriented Architecture (SOA) has been adopted in diverse circulated systems such as World Wide Web services, grid computing schemes, utility computing systems and cloud computing schemes. These schemes are called as Service Oriented Systems. One of the open issues is to prioritize service requests in dynamically altering environments where concurrent instances of processes may compete for assets. If the Cloud Service Provider(CSP) want to prioritize the request, CSP need to monitor the assets that the cloud services have and founded on the available assets the demanded assets can be assigned to the user. Hence, in this paper, propose an approach to find present status of the system by utilizing Dynamic Adaptation Approach. The major target of the research work is to prioritize the service demand, which maximizes the asset utilization in an effective kind that decreases the penalty function for the delayed service. The main concerns should be allotted to requests founded on promise violations of SLA objectives. While most existing work in the area of quality of service supervising and SLA modeling focuses normally on purely mechanical schemes, CSP consider service-oriented systems spanning both programs founded services and human actors. This approach deals with these challenges and assigns priority to the requested service to avoid service delay using Prioritization Algorithm.

Keywords

Cloud Computing, SLA, Resource Allocation, CSP, SOA.

1. INTRODUCTION

Cloud computing is a form of performing IT services in which resources are retrieved through world wide web based tools and submissions rather than a direct attachment to the server. The server contains the data and programs packages that are needed for the users to work remotely. Cloud Computing is about accessing resource that can be always service-based. In cloud environments, the user is able to get access only as services that they required to use and based on their use the cloud can be vary. It is furthermore called as Pay-as-You-Go model, the customer has to pay as they utilize the resources as services. The resources of the cloud can be accessed at anywhere, at any time in the world. Also the cloud has some legal agreement between the Cloud Service Provider and the user is called as Service Level Agreemen.t (SLA). The services can be classified as SaaS (Software as a Service) which the applications in the Internet can be offered as Service, PaaS (Platform as a Service) which provides platform to test, design and test the applications, IaaS (Infrastructure as a Service) which provides storage for servers, storage systems, datacenter.

2. DEPLOYMENT MODEL

The deployment models can be classified a**s**

2.1 Private Cloud

The cloud infrastructure is provisioned for exclusive use by a lone association comprising multiple users (e.g., business units). It may be belongs to, organized, and operated by the organization, a third party, or some blend of them, and it may exist on or off premises.

2.2 Public Cloud

The cloud infrastructure is provisioned for open use by the general public. It may be belongs to, organized, and functioned by a enterprise, learned, or government association, or some blend of them. It lives on the building of the cloud provider.

2.3 Hybrid Cloud

The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that stay unique entities, but are bound simultaneously by normalized or proprietary technology that enables data and application portability (e.g., cloud bursting for burden balancing between clouds).

2.4 Community Cloud

The cloud infrastructure is provisioned for exclusive use by a exact community of buyers from associations that have distributed concerns (e.g., operation, security requirements, principle, and compliance considerations). It may be owned, organized, and functioned by one or more of the organizations in the community, a third party, or some blend of them, and it may exist on or off premises.

3. CHARACTERISTICS

3.1 Resource Pooling

[1]The provider's computing assets are combined to serve multiple consumers utilizing a multi-tenant form, with distinct personal and virtual resources dynamically allotted and re-allotted according to user demand. There is a sense of location self-reliance in that the user generally has no control or information over the accurate location of the supplied resources but may be adapt to specify location at a higher grade of abstraction.

3.2 Rapid Elasticity

[1]Capabilities can be elastically provisioned and issued, in some cases mechanically, to scale quickly outward and inward

commensurate with demand. To the buyer, the capabilities accessible for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

3.3 Measured Service

[1]Cloud systems automatically command and optimize asset use by leveraging a metering capability at some grade of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, User accounts). Asset usage can be supervised, controlled, and reported, supplying transparency for both the provider and buyer of the utilized service.

3.4 Broad network access

[1]Abilities are accessible over the network and that might be gained mechanisms to through components that advertise use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

4. ALLOCATION ISSUES IN CLOUD

The resource allocation problems are placement of virtual machine in datacenters, managing resources of multiple requests of single user, main issue is to find out the status of available resources, can't able to easily transfer huge amount of stored data from one service provider to other service provider, can't able to control over the user resources from remote servers.

5. PROBLEM STATEMENT

Though the cloud has various kinds of resources that has to be allocated instantly to user based on their request. The cloud Service Provider (CSP) doesn't violate the SLA while allocating the resources to the requested user. SLA is an agreement between the CSP and Client about what resources at which time, how and when it can be allocated to the requested client. For that the CSP has to know the present status about the cloud, what are the resources that are available in the cloud, how it can be allocated when multiple clients can be requested for the same resources. The main issues is to avoid the service delay due to some delay that can be unexpected (eg., Traffic in Network) and quality of service (QoS) can be improved.

6. RELATED WORK

Some of the related works about the allocation of resources in the cloud environment are explained as follows.

In [2] the users can enter the cloud at anytime, anywhere to work with their applications and leave the system at anytime when they complete their work. While having this mechanism in cloud, the Cloud Service Provider is able to manage and allocate the resources related to the user requirements and their applications. Having multiple customers, CSP has to satisfy the end-users by efficiently allotting the resources. For that, the job requests can be characterized by the arrival times and teardown times also by the profile of the requirements they need at the activity period. Algorithms can be used to allocate the resources based on the time-variant jobs. Profile Matching and Gap Filling Algorithm achieves maximum efficient allocation of resources in the cloud environment.

In [3] they not only mainly focus on the allocation of resources to real time jobs that can be done before their deadlines but also minimize the cost for the cloud environment they proposed a polynomial-time solution for efficient allocation and the variation of cost while distributing the tasks. And also compared the cost and performance of polynomial-time solution with the optimal solution and Earliest Deadline First (EDF) method. Based on the user requirements, user can select distinct types of computing resources. In user application, it has a set of tasks, each task has its arrival time and deadline. If the tasks doesn't have enough Virtual Machines(VM's) to complete, it searches for the cheapest VM to complete the tasks by using the lookup table that has a range of computing speeds from different types of VM's. It uses EDF Greedy Algorithm that allocates the task to VM's.

In [4] the author focus on IaaS is how the VM utilize the resources that satisfy the Quality of Service (QoS) and minimizing the operating costs. The main issue is to migration of VM to Physical nodes and dynamic allocation of resources to VM's. The Author proposed a two-tier resource manager with a utility function for allocating the resources dynamically to VM's by local controller and global controller maximizes the local node utility function for migrating the load shares between the VM's.

7. PROPOSED WORK

To overcome this problem, priorities should be assigned to incoming request based on the potential of SLA objectives and balancing the load to avoid delay of system in the overloaded system. To improve allocation and utilization of resources dynamically for the clients/users, cloud distributes workloads in the overloaded system across multiple computing resources that reduces the overload of the system. If the client may be the old customer the first priority assigned to them and then assigns priority for the new customers. For prioritizing the request, CSP uses the algorithm that states the execution state of processes which are in accessible in service-oriented systems and for the delay of service the penalty functions are provided that based on the SLA's.

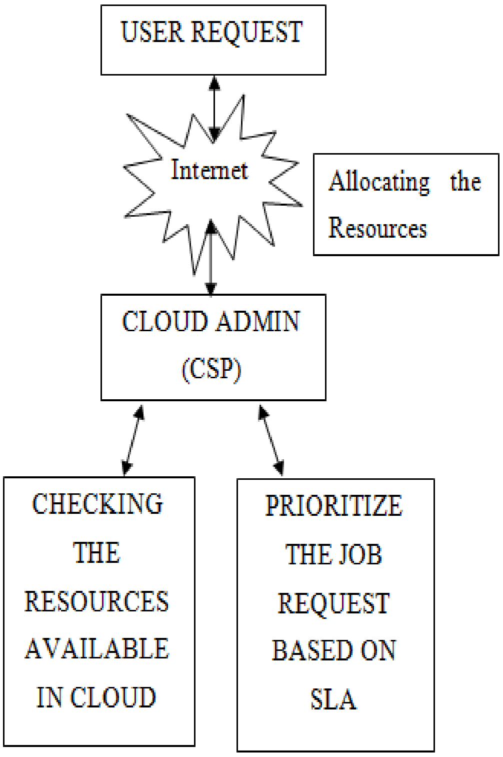
USER REQUEST
Internet
Allocating the Resources
CLOUD ADMIN (CSP)
CHECKING THE RESOURCES AVAILABLE IN CLOUD
PRIORITIZE THE JOB REQUEST BASED ON SLA

Figure 1. Resource Allocation in Cloud

We use prioritize algorithm for ordering the job requests from the client/cloud users based on the SLA. CSP also assures about the reduction of penalty for the delayed service and to improve the quality of service they provide for the client.

8. PRIORITIZATION ALGORITHM

```
INPUT: Service S, Response Time $S_{RT}$, a set of processors P, Penalty Function for each Process $L_P(t)$
OUTPUT: Ranked Users Job Requets
For each process p in P do
$R_p$= Pending user job requests of S in p
$R_e$=user job requests of S predicted to be made during $S_{RT}/2$ period in each process p
Assume all user job requests replies in R are received after $S_{RT}$, predict time t of finish for each process p.
$l_0=L_p(t)$  //Default Penalty for Each Process
for each request r in R do
Assume that a reply of r is received after $S_{RT}*2$ and all other user job requests replies in Rare received after $S_{RT}$, predict time $t_r$ of finish for p
$l_r=L_p(t_r)$  //Current user job request Penalty
$d_r=l_r-l_0$  //Difference between the default penalty and current user job request penalty

In List D, add the tuple r, $d_r$ and request
time k
End
End
Sort D for descending by $d_r$ and then ascending by k
Return D
```

9. METHODOLOGY

The contribution of work for allocating the resources as follows:

The client is the first module who requests for the service from the cloud service provider. Client/Customer only the main objective who gets service from the cloud service provider so that CSP gain profit by renting the services to clients. This phase is based on the service request needed for their application. The services can be requested based on some agreements that has to be satisfy by both client and CSP.

The second is about the monitoring of present status of the cloud environment by using dynamic adaptation approach. It's a runtime approach have their own paths that support problems and resolves for complex resources. So that the presence of resource availability and demand on the resources can be identified. The resources can be scheduled and allotted based on the SLA, that doesn't violate it.

The third phase is about the availability of the resources that can be found by the CSP/Cloud Admin. Based on the availability of resources in the cloud the new job requests can be accepted. The CSP is responsible for the allocating the resources for the requested client. Also the CSP focus on the Quality of Service (QoS) and any service delay that leads to the penalty for delayed service.

Lastly based on the SLA and service request, the resources can be prioritized using the Prioritization Algorithm that reduces penalty for delayed service and improves the quality of service. The prioritized requests doesn't violate the Service Level Agreements (SLA's) . It get the inputs such as Service, arrivals, deadline and its penalty function. For each process calculate the pending request and calculate the penalty of current job request then after return to list. The delay of service can be reduced and QoS can be improved.

10. CONCLUSION

The problem of delaying of service requested by the user of unexpected delay is focused in this paper. The Model for scheduling the request in service-oriented systems and prioritization algorithm can be proposed. The proposed solution results in reduction of service delay in the cloud environment that shows the advantage. The CSP assigns the priority for user service request that avoids the violation of SLA objectives.

The future work is to improve the QoS and to handle the request from different sources at multiple clients. And also to reduce the burden of allocating the resources, an idea of creating the instance that can support for the CSP to handle different service requests at any instances of time. This may compete the CSP to get more response from the clients and improves the efficiency and effectiveness of the resources available. The main idea behind this is to avoid the unexpected delays in the cloud environment so that the penalty functions can be decreased and achieves maximum profit for the Cloud Provider.

REFERENCES

[1] Peter Mell, Timothy Grance,"The NIST definition of Cloud Computing", Recommendations of the National Institute of Standards and Technology, U.S Department of Commerce, 800-145.

[2] Davide Tammaro, A.Doumith, Jean-Pauls Smets, Maurice Gagnaire, Sawsan Al Zahr, "Dynamic Resource Allocation in Cloud Environment Under Time-Variant Job Requests", 3rd IEEE International on Cloud Computing Technology and Science, 978-0-7695-4622-3/11

[3] Karthik Kumar, Jing Feng, Yamini Nimmagada, Yung-Hsiang Lu, “Resource Allocation for Real-Time Tasks using Cloud Computing”, IEEE, 2011, 978-1-4577-0638-7 /11

[4] Bernd Freisleben, Dorian Minarolli, “Utility-based Resource Allocation for Virtual Machines in Cloud Computing”, IEEE, 2011, 978-1-4577-0681-3/11

[5] Roman Khazankin, Daniel Schall, Schahram Dustdar Adaptive Request Prioritization in Dynamic Service-oriented Systems, 2011 IEEE International Conference on Services Computing, 978-0-7695-4462-5/11

[6] J.Lakshmi, Mohit Dhingra, S.K. Nandy (2012), “Resource Usage Monitoring in Clouds”, IEEE.

[7] Akshatha M, K C Gouda, Radhika T V (2013), “Priority Based Resource Allocation Model for Cloud Computing”,International Journal of Science, Engineering and Technology Research, Vol 2, Issue 1.

[8] Jaisankar N, Sendhil Kumar KS, Vignesh V (2013), “Resource Management and Scheduling in Cloud Environment”, International Journal of Scientific and Research Publications, Vol 3, Issue 6.

[9] M. Armbrust et al., “Above the Clouds: A Berkeley View of Cloud Computing,” technical report, Univ. of California, Berkeley, Feb.2009.

[10] L. Siegele, “Let It Rise: A Special Report on Corporate IT,” The Economist, vol. 389, pp. 3-16, Oct. 2008.

[11]] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, “Xen and the Art of Virtualization,” Proc. ACM Symp. Operating Systems Principles (SOSP ’03), Oct. 2003.

[12] “Amazon elastic compute cloud (Amazon EC2),” http://aws. amazon.com/ec2/, 2012.

[13] C. Clark, K. Fraser, S. Hand, J.G. Hansen, E. Jul, C. Limpach, I.Pratt, and A. Warfield, “Live Migration of Virtual Machines,” Proc. Symp. Networked Systems Design and Implementation (NSDI ’05), May 2005.

[14] A. Chandra, W. Gongt and P. Shenoy. Dynamic resource allocation for shared clusters using online measurements. International Conference on Measurement and Modeling of Computer Systems SIGMETRICS 2003.

[15] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat and R.P.Doyle. Managing energy and server resources in hosting centers. Presented at 18th ACM Symposium on Operating Systems Principles (SOSP'01), October 21, 2001.

[16] M. N. Bennani and D. A. Menasce. Resource allocation for autonomic clusters using analytic performance models. Presented at Second International Conference on Autonomic Computing.