RESEARCH ARTICLE OPEN ACCESS

Heterogeneous Resource Allocation Strategies in Elastic Cloud

Neha Solanki
(Computer Science Department, Jai Narain Vyas University, and Jodhpur)

Abstract:

Cloud computing deals with computation, software, data access and storage services. The main goal of cloud computing is to make a better use of distributed resources, combine them to achieve higher throughput and be able to solve large scale computation problem. In elastic cloud, user's demand changes dynamically so proper resource allocation and scheduling is necessary. Efficient and effective resource allocation and scheduling are the basis for excellent performance of clouds, because all the quality of service constraints, like throughput, maximum efficiency, response time and power consumption are mainly dependent on the mechanism of heterogeneous resource allocation and scheduling.

Keywords — **Resource Allocation, Energy Efficiency, Virtual Machines, Best-Fit Algorithm**

I. INTRODUCTION

Cloud computing [1, 2] has recently received attention as a new computing paradigm to provide dynamically scalable and virtualized resource as a service over the Internet. By this means, users are able to access the resources, such as applications and data, from the cloud anywhere and anytime on demand. Currently, several large companies, such as Amazon, Google, Yahoo!, Microsoft, IBM, and Sun have developed their own cloud platforms for consumers and enterprises to access the cloud resources through services.

The definition proposed by the National Institute of Standards and Technology, USA [3] is "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." The origin of the term comes from the early days of the Internet where the network was depicted as a cloud (Fig. 1.1).

II. RESOURCE ALLOCATION AND SCHEDULING IN CLOUDS

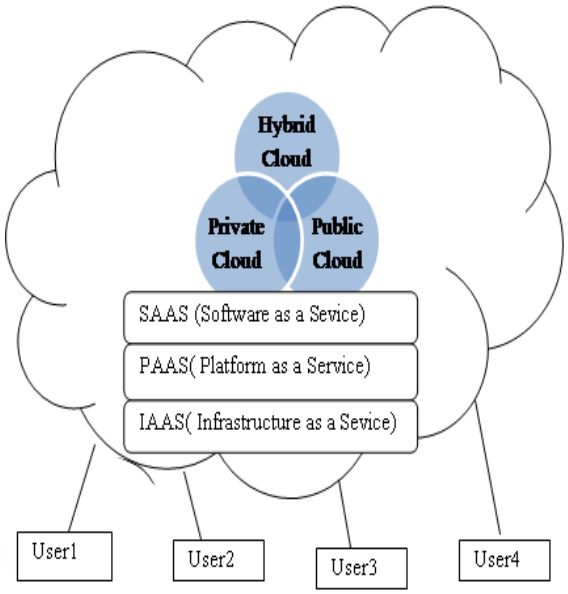

Figure 1.1: The Cloud Computing Model

In cloud computing various cloud consumers demand variety of services as per their dynamically changing needs. So it is the job of cloud computing to avail all the demanded services to the cloud consumers. All the Quality of Services (QoS) constraints, like throughput, response time and power consumption, are mainly dependent on the mechanism of resource allocation and scheduling, so these mechanism should be done in efficient and

effective way. The cloud computing is the concept of provide the virtualized resources to the multiple user's at a time. Hence the cloud computing environment required a powerful resource allocation and scheduling mechanism to make it reliable. But due to the availability of finite resources it is very difficult for cloud providers to provide all the demanded services in time. From the cloud providers' perspective cloud resources must be allocated in a fair manner.

The ultimate goal of resource allocation and scheduling in cloud computing is to maximize the profit for cloud providers and to minimize the cost for cloud consumers. Key issues in resource allocation and scheduling of clouds are:

a) Quality of Service: Resource allocation and scheduling should be done in such a manner that it maintains the quality of service.
b) Energy-Efficient clouds: A large-scale computing infrastructure consumes enormous amounts of electrical power leading to operational costs that exceed the cost of the infrastructure in few years. For example, in 2006 the cost of electricity consumed by IT infrastructures in US was estimated 4.5 billion dollars and tends to double by 2011 [4].
c) Heterogeneous Environment: Many cloud applications largely assume a homogeneous environment. For example, Hadoop [5] assumes that all nodes participating in the cluster have the same processing power. To take a full advantage of available hardware, cloud-oriented applications must be heterogeneous-aware.
d) Unpredictability: the cloud environment is highly variable and unpredictable. To increase resource utilization, providers try to oversubscribe as many users to a shared infrastructure. This results in resource contention and interference. Other factors that contribute to unpredictability of the environment include heterogeneity within the same instance type and administrative action to maintain the service level. These make it extremely difficult to predict the performance variability and track down its causes.

III. RESOURCE ALLOCATION STRATEGIES IN CLOUDS

Resource allocation is the process which involves deciding how many, where, and when to allocate the resource to a particular or set of virtual machines. Resource allocation is one of the most challenging problems in high performance computing field for managing and provides effective utilization of resources. All the quality of services (QoS) constraints, like throughput, maximum efficiency, response time and power consumption are mainly dependent on the mechanism of efficient and effective resource utilization. The effective resource utilization is directly derived from fine grained efficient resource allocation mechanism. The cloud computing is the concept of provide the virtualized resources to the multiple user's at a time. Hence the cloud computing environment is also required a powerful resource mechanism to make it reliable. The efficient resource allocation and proper scheduling make it possible to distribute recourse appropriately. Fig 1.2 shows the cloud computing architecture for resource allocation.

Four Types of resource allocation strategies; they are Best Fit/First Fit Resource Allocation, Genetic Algorithm with Multiple Fitness Resource Allocation, Energy Efficient Resource Allocation, SLA-based Resource Allocation [6, 7, 8, 9]. Best Fit/First Fit resource allocation and Genetic algorithm with multiple fitness resource allocation, these two strategies focus on load balancing (RAM, CPU and Bandwidth). SLA-based resource allocation focuses on Quality of Service (QoS), obligations, and penalties in case of agreement violations. While energy efficient resource allocation focuses on both load balancing and Quality of Service (QoS). Energy efficient resource allocation is highly scalable strategy. Best Fit/First Fit resource allocation, Genetic algorithm with multiple fitness resource allocation and energy efficient resource allocation includes bandwidth in there load balancing strategies while SLA-based resource allocation does not.

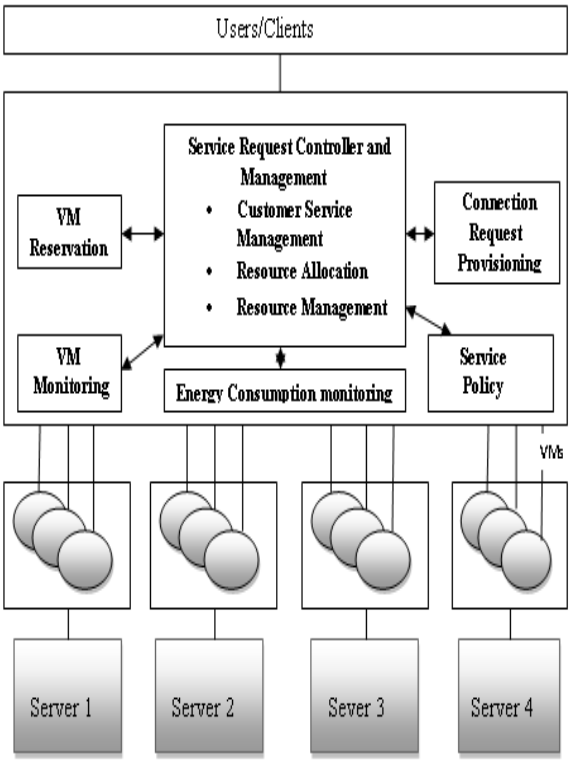

Fig 1.2: Cloud Computing Architecture for Resource Allocation

TABLE I
COMPARISONS OF RESOURCE ALLOCATION ALGORITHMS

Features	Best Fit/First Fit Resource Allocation	Genetic Algorithm with Multiple Fitness Resource Allocation	Energy Efficient Resource Allocation	SLA-based Resource Allocation
SLA(QoS)	Low	Low	Medium	High
Energy Conservation	Medium	Medium	High	Low
Optimum Resource Allocation	High	High	High	Medium
Load balancing	High	High	High	Low
Scalability	Low	Low	High	Medium
Bandwidth	High	High	High	Low
Application Dependent	Low	Low	Low	High

SLA-based resource allocation is application dependent while the others are not. Best Fit/First Fit resource allocation, Genetic algorithm with multiple fitness resource allocation and energy efficient resource allocation optimizes the resource allocation while SLA-based resource allocation does not.

IV. CONCLUSIONS

Reliable, scalable and inexpensive dynamic computing environments for end-users is a key challenge to service providers, which can be solve by efficient and effective resource allocation and scheduling. Strategies for resource allocation are all about integrating cloud provider activities to utilize and allocate scarce resources within the limit of cloud environments to meet the needs of the cloud applications. Table I shows the comparisons of these four resource allocation algorithms.

REFERENCES

1. *M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.*
2. *R. Buyya, C. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in The 10th IEEE international conference on high performance computing and communications. IEEE, 2008, pp. 5–13.*
3. *Peter Mell and Timothy Grance, The NIST Definition of Cloud Computing, National Institute of Standards and Technology, USA, Special Publication 800-145, September 2011.*
4. *R. Brown et al., "Report to congress on server and data center energy efficiency: Public law 109-431," Lawrence Berkeley National Laboratory, 2008.*
5. *Hadoop. http://hadoop.apache.org/core/*

6. *T. C. Ferreto, M. A. S. Netto and et al. Server consolidation with migration control for virtualized data centers. Future Generation Comp. Syst. 2011, 27(8):1027-1034.*

7. *Efficient Resource Allocation in Cloud Data Centers Through Genetic Algorithm. Ehsan Arianyan, Davood maleki, Alireza Yari, Iman Arianyan. 6'th International Symposium on Telecommunications (IST'2012).*

8. *D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," Cluster Computing,vol. 12, no. 1, pp. 1–15, 2009.*

9. *Vincent C. Emeakaroha, Ivona Brandic, Michael Maurer, Ivan Breskovic,"SLA-Aware Application Deployment and Resource Allocation in Clouds", 35th IEEE Annual Computer Software and Applications Conference Workshops (2011).*