



A REVIEW ON RESOURCE ALLOCATION MECHANISM IN CLOUD ENVIRONMENT

Rajveer Kaur ⁽¹⁾, Mr. Navtej Singh Ghumman ⁽²⁾

⁽¹⁾ Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
rajveersaini.17@gmail.com

⁽²⁾ Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
navtejghumman@yahoo.com

ABSTRACT

Cloud computing is Internet based development and use of computer technology. It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them. Cloud computing is a hot topic all over the world nowadays, through which customers can access information and computer power via a web browser. As the adoption and deployment of cloud computing increase, it is critical to evaluate the performance of cloud environments. Currently, modeling and simulation technology has become a useful and powerful tool in cloud computing research community to deal with these issues. Cloud simulators are required for cloud system testing to decrease the complexity and separate quality concerns. Cloud computing means saving and accessing the data over the internet instead of local storage. In this paper, we have provided a short review on the types, models and architecture of the cloud environment.

Keywords

Cloud Computing, Virtual Machine, Data Center, Data Center Broker, Load Balancing.

INTRODUCTION

Cloud computing is an internet-based computing technology, where shared resources such as software, platform, storage and information are provided to customers on demand. Cloud computing is a computing platform for sharing resources that include infrastructures, software, applications, and business processes. Cloud Computing is a virtual pool of computing resources. It provides computing resources in the pool for users through internet. Cloud computing, as an emerging computing paradigm aiming to share storage, computation, and services transparently among a massive users. The exact definition of cloud computing is A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet [5]. Current cloud computing systems pose serious limitation to protecting users data confidentiality. Since users sensitive data is presented in unencrypted forms to remote machines owned and operated by third party service providers, the risks of unauthorized disclosure of the users sensitive data by service providers may be quite high. There are many techniques for protecting users data from outside attackers. An approach is presented to protecting the confidentiality of users data from service providers, and ensures service providers cannot collect users confidential data while the data is processed and stored in cloud computing systems. Cloud computing systems provide various Internet based data storage and services. Due to its many major benefits, including cost effectiveness and high scalability and exibility, cloud computing is gaining significant momentum recently as a new paradigm of distributed computing for various applications, especially for business applications. Along with the rapid growth of the Internet. With the rise of the era of cloud computing, concerns about Internet Security continue to increase. To address this problem we propose the design of a system that will capture the movement of information on the cloud. We will be identifying whether there is a need for some type of security capture device/measure on the cloud, which will allow users to know whether their information is secure and safe without comprising from threats and attacks.

EVOLUTION OF CLOUD COMPUTING

Cloud computing began to get both awareness and popularity in the early 2000s. When the concept of cloud computing originally came to prominence most people did not fully understand what role it fulfilled or how it helped an organization. In some cases people still do not fully understand the concept of cloud computing. Cloud computing can refer to business intelligence (BI), complex event processing (CEP), service-oriented architecture (SOA), Software as a Service (SaaS), Web-oriented architecture (WOA), and even Enterprise 2.0. With the advent and growing acceptance of cloud-based applications like Gmail, Google Calendar, Flickr, Google Docs, and Delicious, more and more individuals are now open to using a cloud computing environment than ever before. As this need has continued to grow so has the support and surrounding infrastructure needed to support it. To meet those needs companies like Google, Microsoft, and Amazon have started growing server farms in order to provide companies with the ability to store, process, and retrieve data while generating income for themselves. To meet this need Google has brought on-line more than a million servers in over 30 data centers across its global network. Microsoft is also investing billions to grow its own cloud infrastructure. Microsoft is currently adding an estimated 20,000 servers a month. With this amount of process, storage and computing power coming online, the concept of cloud computing is more of a reality than ever before. The growth of cloud computing had the net effect of businesses migrating to a new way of managing their data infrastructure. This growth of cloud computing capabilities has been described as driving massive centralization at its deep center to take advantage of economies of scale in computing power, energy consumption, cooling, and administration.

CLOUD ARCHITECTURE

The architecture of cloud involves multiple cloud components communicating with each other over the application programming interfaces (APIs), usually web services. The two most significant components of cloud computing architecture are known as the front end and the back end. The front end is the part seen by the client, i.e. the customer. This includes the clients network or computer, and the applications used to access the cloud via a user interface such as a web browser. The back end of the cloud computing architecture is the cloud itself, which comprises of various computers, servers and data storage devices. Cloud services may be offered in various forms from the bottom layer to top layer in which each layer represent one service model. The three key cloud delivery models are software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). Infrastructure-as-a-Service (IaaS) is offered in the bottom layer, where resources are aggregated and managed physically (e.g., Emulab) or virtually (e.g., Amazon EC2), and services are delivered in forms of storage (e.g., GoogleFS), network (e.g., Open ow), or computational capability (e.g., Hadoop MapReduce). The middle layer delivers Platform-as a-Service (PaaS), in which services are provided as an environment for programming (e.g., Django) or software execution (e.g., Google App Engine). Software- as-a Service (SaaS) locates in the top layer, in which a cloud provider further confines client flexibility by merely offering software applications as a service. Apart from the service provisioning, the cloud provider maintains a suite of management tools and facilities (e.g., service instance life-cycle management, metering and billing, dynamic configuration) in order to manage a large cloud system.

TYPES OF CLOUD

Cloud computing is typically classified in two ways [9]:

1. Location of the cloud computing

2. Type of services offered

Location of the cloud

Cloud computing is typically classified in the following ways:

1. Public cloud: In Public cloud the computing infrastructure is hosted by the cloud vendor at the vendor's premises. The customer has no visibility and control over where the computing infrastructure is hosted. The computing infrastructure is shared between any organizations.

2. Private cloud: The computing infrastructure is dedicated to a particular organization and not shared with other organizations. Some experts consider that private clouds are not real examples of cloud computing. Private clouds are more expensive and more secure when compared to public clouds.

Private clouds are of two types: On-premise private clouds and externally hosted private clouds. Externally hosted private clouds are also exclusively used by one organization, but are hosted by a third party specializing in cloud infrastructure. Externally hosted private clouds are cheaper than On-premise private clouds.

3. Hybrid cloud: Organizations may host critical applications on private clouds and applications with relatively less security concerns on the public cloud. The usage of both private and public clouds together is called hybrid cloud. A related term is Cloud Bursting. In Cloud bursting organization use their own computing infrastructure for normal usage, but access the cloud for high/peak load requirements. This ensures that a sudden increase in computing requirement is handled gracefully.

4. Community cloud involves sharing of computing infrastructure in between organizations of the same community. For example all Government organizations within the state of California may share computing infrastructure on the cloud to manage data related to citizens residing in California.

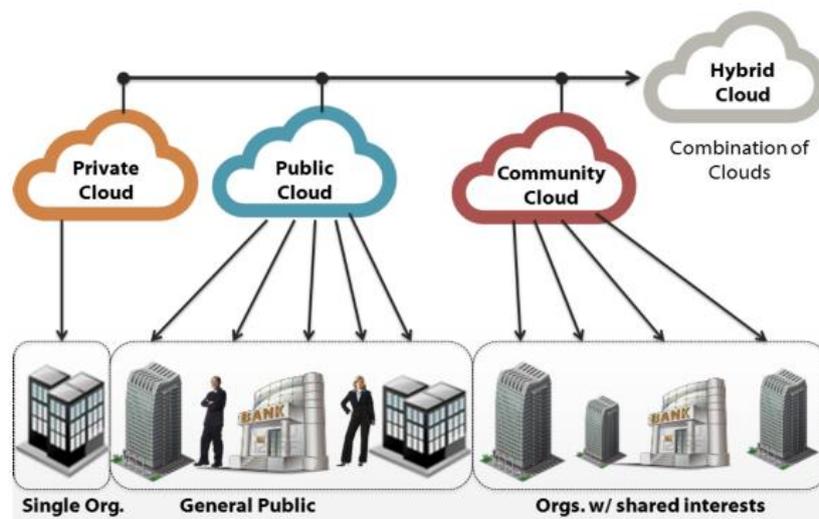


Figure 1. Types of Cloud

CLASSIFICATION BASED UPON SERVICES PROVIDED

Based upon the services offered, clouds are classified in the following ways:

- 1. Infrastructure as a service (IaaS)** involves offering hardware related services using the principles of cloud computing. These could include some kind of storage services (database or disk storage) or virtual servers. Leading vendors that provide Infrastructure as a service are Amazon EC2, Amazon S3, Rackspace Cloud Servers and Flexiscale.
- 2. Platform as a Service (PaaS)** involves offering a development platform on the cloud. Platforms provided by different vendors are typically not compatible. Typical players in PaaS are Google's Application Engine, Microsofts Azure, Salesforce.com's force.com .
- 3. Software as a service (SaaS)** includes a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector. The pioneer in this field has been Salesforce.coms offering in the online Customer Relationship Management (CRM) space. Other examples are online email providers like Googles gmail and Microsofts hotmail, Google docs and Microsofts online version of office called BPOS (Business Productivity Online Standard Suite).

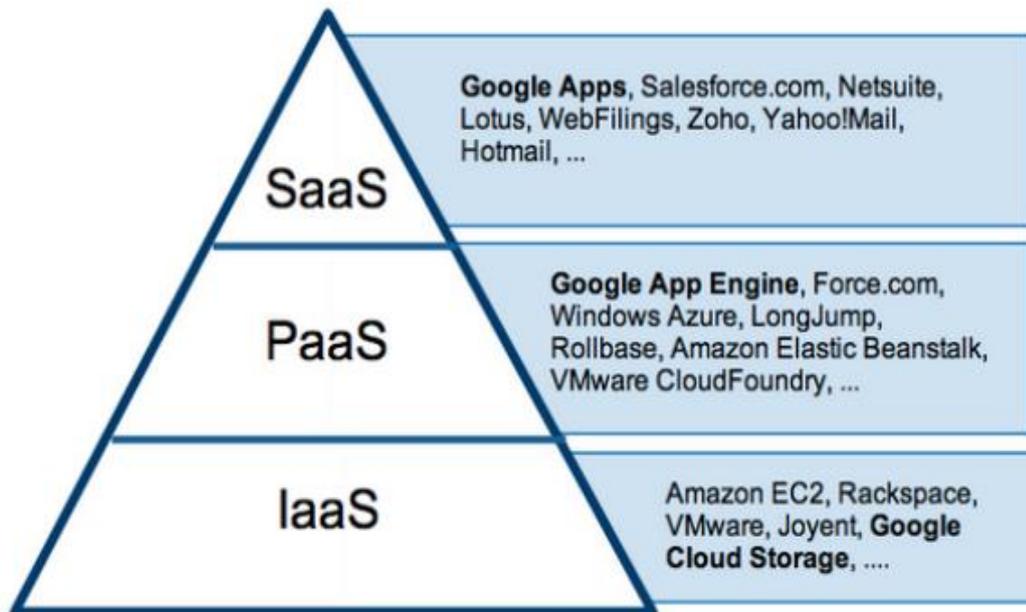


Figure 2. Cloud delivery model

BENEFITS OF CLOUD COMPUTING

Growth of Cloud Computing is huge with respect to personal and business uses. Cloud users can access the online resources. There are some benefits of cloud computing :-

- 1. Scalability:** Scalability is the capability of a system to handle the growing amount of task in an elegant manner and its ability to enhance total throughput when resources are added. Resources can be hardware, servers and storage. The user can increase or decrease the resources according to their requirement without buying the resources.
- 2. Mobility:** Mobility means users can operate the applications from anywhere, anytime over the internet. Cloud computing supports the mobility.
- 3. Low Infrastructure Costs:** Cloud computing supports the pay-per-use model and helps an organization to pay for the resources they need, no need to pay for the resources that are available in the cloud.
- 4. Reduce capital costs:** We don't require spending money on hardware/software.
- 5. Enhance accessibility:** Data can be accessed anytime, anywhere through the internet. Thus, the cloud computing offers many benefits in the form of elasticity, availability, increased storage, reduce cost and expandability on-demand but there are some limitations in cloud computing that are:
 1. Requires a constant Internet connection: Cloud computing is dependence on network connectivity. It is impossible if you cannot connect to the Internet.
 2. Slow speed of Interment: Uploading and downloading of huge documents may take a long time.

LOAD BALANCING

Load balancing is one of the main issues related to cloud computing. The load can be a memory, CPU capacity, network or delay load. It is always required to share work load among the various nodes of the distributed system to improve the



resource utilization and for better performance of the system. This can help to avoid the situation where nodes are either heavily loaded or under loaded in the network. Load balancing is the process of ensuring the evenly distribution of work load on the pool of system node or processor so that without disturbing, the running task is completed. The goals of load balancing [16] are to:

- Improve the performance
- Maintain system stability
- Build fault tolerance system
- Accommodate future modification.

There are mainly two types of load balancing algorithms:

STATIC ALGORITHM In static algorithm the traffic is divided evenly among the servers. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static algorithm is proper in the system which has low variation in load.

DYNAMIC ALGORITHM

In dynamic algorithm the lightest server in the whole network system is searched and preferred for balancing a load. For this real time communication with network is needed which can increase the traffic in the system. Here current state of the system is used to make decisions to manage the load.

CLOUD SIM

The CloudSim simulation layer provides support for modeling and simulation of virtualized Cloud-based data center environments including dedicated management interfaces for VMs, memory, storage, and bandwidth. The fundamental issues, such as provisioning of hosts to VMs, managing application execution, and monitoring dynamic system state, are handled by this layer. A Cloud provider, who wants to study the efficiency of different policies in allocating its hosts to VMs (VM provisioning), would need to implement his strategies at this layer. Such implementation can be done by programmatically extending the core VM provisioning functionality. There is a clear distinction at this layer related to provisioning of hosts to VMs. A Cloud host can be concurrently allocated to a set of VMs that execute applications based on SaaS provider's defined QoS levels. This layer also exposes the functionalities that a Cloud application developer can extend to perform complex workload profiling and application performance study. The top-most layer in the CloudSim stack is the User Code that exposes basic entities for hosts (number of machines, their specification, and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies. By extending the basic entities given at this layer, a Cloud application developer can perform the following activities: (i) generate a mix of workload request distributions, application configurations; (ii) model Cloud availability scenarios and perform robust tests based on the custom configurations; and (iii) implement custom application provisioning techniques for clouds and their federation.

RESEARCH MOTIVATION

Cloud computing thus involving distributed technologies to satisfy a variety of applications and user needs. Sharing resources, software, information via internet are the main functions of cloud computing with an objective to reduced capital and operational cost, better performance in terms of response time and data processing time, maintain the system stability and to accommodate future modification in the system. So there are various technical challenges that needs to be addressed like Virtual machine migration [20], server consolidation, fault tolerance, high availability and scalability but central issue is the load balancing, it is the mechanism of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Load Balancing [5] is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) [32] customized for a specific use. They have the ability to handle the high speed network traffic whereas Software-based load balancers run on standard operating systems and standard hardware components. Clients request for the virtual machine and cloud broker handles the client request according to available virtual machine. If the VM is idle then broker allocate that VM to user for the processing and if VM is not free then incoming requests of the client are send into the waiting state until the resources are free.

Most of the existing load balancing algorithms will work only in the homogeneous cloud system where all the resources are of same configuration. Cloudlets or the tasks of the users are not assigned to the virtual machine according to their capacities. There may be a scenario where a cloudlet with highest priority is assigned to the machine with lowest capacity in the host. The processing capacity (No of processors / MIPS) is not considered for assigning the VM to a job. The status or the current load of the VM is not computed before allocating the new request to it. There are several parameters that should be checked before allocating the new request to the virtual machine.



CONCLUSION

This paper discusses a new wave of online services in the field of information technology: cloud computing with its challenges. In cloud computing, there are infinite computing capabilities with attractive pay-per-use scheme. Cloud computing provide everything as a service to their users, like as: storage of data as a service, application software as a service, computing platform as a service and computing infrastructure as a service etc. However this wave still needs to resolve some of its existing issues with urgency. One of the major issues of cloud computing is system load balancing, because overloading of a particular node makes it slow down resulting poor system efficiency. So there is always a requirement of efficient load balancing algorithms for improving the utilization of computing resource.

REFERENCES

- [1] O.M. Elzeki . "Improved Max-Min Algorithm in Cloud Computing". International Journal of Computer Applications(0975-8887) Volume 50-No.12,july 2012.
- [2] "A technical support seminar on cloud computing technology" by Prashant Gupta.
- [3] Amandeep Kaur Sidhu. "Analysis of load balancing techniques in cloud computing". International Journal of computers & technology volume 4 No. 2, March-April, 2013, ISSN 2277-3061.
- [4] Ektemal Al-Rayis. "Performance Analysis of load balancing Architectures in Cloud computing" 2013 European Modeling Symposium. 978-1-4799-2578-0/13\$31.00@2013 IEEE.
- [5] Haozheng Ren. "The load balancing Algorithm in cloud computing Environment" 2nd International Conference on computer science and network technology 2012.
- [6] Tushar Desai. "A survey of various load balancing techniques and challenges in cloud computing" International Journals of scientific and technology research volume 2. Issue11,Nov2013.
- [7] Upendra Bhoi. "Enhanced max-min Task scheduling Algorithm in cloud computing". International Journal of Application or Innovation in Engineering & management(IJAIEEM), April 2013.
- [8] Klaithem Al Nuaimi, "A survey of load balancing in cloud computing challenges and algorithm". 2012 IEEE second symposium on network cloud computing and applications.
- [9] Gytis Vilutis, "Model of load balancing and scheduling in cloud computing". Proceedings of the ITI 2012 34th Int.Conf. on Information Technology Interfaces, June 25-28,Cavat,Croatia.
- [10] S. Banerjee, I. Mukherjee and P.K. Mahanti, Cloud Computing Initiative using Modified Ant Colony Framework, World Academy of Science and Technology, 56, pp. 221-224, 2009.