Academic Guru Publication House

Advances in Cloud Computing Security
Techniques and Applications
Volume 1
Year: 2021

AG PH Books

An Analysis of Cloud Computing's Resource Allocation Methods

A. Shenbaga Bharatha Priya[1*]

[1]*Teaching Fellow, Ramanujan Computing Centre, Anna University*

Abstract

In order for a cloud subscriber to run their applications on many platforms, anywhere, at any time, utilising just the resources given by the cloud service provider, the cloud service providers play a critical role. As a result, cloud customers must still deal with the challenge of quickly accessing the computing resources they need. For many users in several applications, this will have an effect on the service time and service level agreements. As a consequence, earlier research on cloud resource allocation need to be completed. This study examines the distribution and management of cloud computing resources. There is a lot of discussion here on alternative resource allocation systems and the difficulties they pose.

Keywords: Cloud computing; Resource Allocation; Infrastructure; service of Clouds.

1. INTRODUCTION

Cloud computing is nothing more than the Internet's mutation. On the cloud, it's possible that people to have all of the resources they need. When it comes to on-demand IT services and products, cloud computing is the natural next step. It's just a matter of time until cloud computing becomes a critical component in the development and deployment of distributed applications. The availability of a wide range of resources in the cloud is helping to make cloud computing more popular among the general public. Many cloud service providers including Microsoft, Amazon, Google and IBM provide platform as a service (PaaS). Interoperability capabilities enabled developers to distribute apps among machines hosted by a single enterprise. A cloud computing provider manages and deploys a broad network of

* ISBN No. 978-81-955340-6-7

computer resources for these applications. While developers benefit from a managed computing platform, they do not have to spend time and money setting up and maintaining their own network infrastructures. As a result, the cloud must address how to manage quality of service (QoS) for cloud users who share resources and establish service level agreements (SLAs). Clients may access all of their resources via a single point of access provided by cloud computing. Cloud computing provides several advantages, including a defined and abstracted infrastructure, a completely virtualized environment, dynamic infrastructure, pay-as-you-use, and no software or hardware installations. As a result, the most pressing concern is the sequence in which the requests are granted. As a consequence, the scheduling of resources is affected. Effective usage of system resources is crucial to getting the best potential performance out of the system. As a result, cloud computing services are mostly marketed on a per-minute or per-hour basis. Because of this, it is important to plan the use of resources effectively. When it comes to cloud platforms, there are two degrees of resource allocation (or load balancing). At the time of upload, a load balancer distributes the requested instances among physical machines in an effort to balance the computational demand of various applications. Multiple requests for the same application should each be assigned to a distinct instance in order to spread out the computational load as evenly as possible. Elastic load balancing (ELB), for example, is used by Amazon EC2 to manage how incoming requests are handled. Requests to specified availability zones, individual instances, or instances with the fastest response times may be directed by application designers. Here, we explore the importance of allocating resources in the following paragraphs.

2. SIGNIFICANCE OF RESOURCE ALLOCATION

RA in cloud computing is an internet-based technique for distributing resources to online cloud applications. It is not necessary for suppliers of service to manage the resources assigned to each individual module as a consequence of the resource provisioning process. The Resource Allocation Method (RAS) is an integrated strategy for utilising and allocating limited resources in the cloud environment to meet cloud application needs. All of the resources required to execute a task requested by a user must be gathered in this manner. The best RAS takes into account the timing and sequence of resource distribution. The following characteristics should be avoided in an ideal RAS:

- There is a problem known as resource disputation when two applications try to use the same resource at the same time.
- There is a high demand for resources when there are a limited number of resources.
- When resources are separated, a scenario known as "resource fragmentation" emerges. [Resources will be available there, however they can't be allocated to the required application.]
- A situation known as "over-provisioning" occurs when an application is provided with more resources than it requested

- Resources are not allocated to an application in sufficient quantity to meet the expected demand.

3. AT A GLANCE: THE MANAGEMENT OF RESOURCES

3.1. Resource Management

In our view, research management encompasses the finding, allocation, and monitoring of resources, as seen in Fig. 1 below. These processes are responsible for managing physical resources such as CPU cores, disc storage, and network bandwidth. Slices of these resources must be allocated to virtual machines executing a wide variety of tasks, all of which might potentially compete for the same resources. Here is a breakdown of the taxonomy for resource management components:

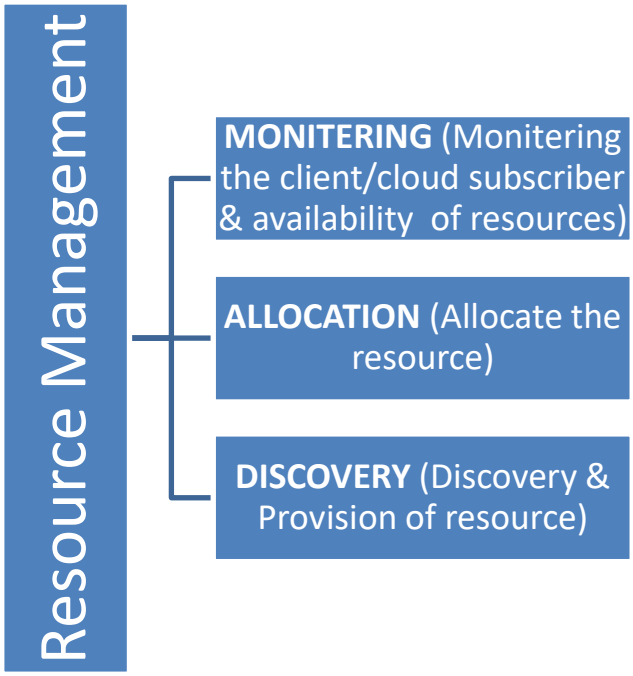

Figure: 1 Elements of resource management.

The discovery process is the core of resource management. Searching for the right resources to meet the application's needs is the first step. [1] In this case, the cloud provider is in charge of overseeing the whole thing. The resource broker or user broker is doing this step in order to find out what resources are accessible. The discovery process includes detailed explanations of the resources accessible. A resource management system (RMS) and other RMSs that interact with it may learn the status of the resources that they manage via resource discovery, according to [2]. The discovery of resources works in

conjunction with the distribution of resources to give the information server with data on the current condition of resources.

Using the Internet, a cloud application may be assigned to a source that is accessible and ready to go. Using a pay-as-you-go model, these resources are distributed to users depending on their requests. Scheduling and dispatching are used to assign resources in this process. The client's given resources will be scheduled by the scheduler. As a result, the dispatcher will assign the client's resources. There are a number of ways to monitor and manage hardware and software infrastructures, including resource monitoring as specified in paper [4]. It also gives data and Key Performance Indicators (KPIs) to assist in the allocation of resources for both platforms and applications in the cloud. When there is a problem with the physical layer or the services layer, this is a key component for monitoring the state of the resources. In our perspective, allocation and discovery should be included in provisioning, but tracking should be managed as a distinct process. All three processes, however, are interrelated in order to offer customers with resources. In the next part, we'll go through how the resource provisioning and resource management processes work.

4. RESOURCE ALLOCATION STRATEGIES & ALGORITHMS

a) Topology Aware Resource Allocation (TARA)

Allocating resources in the cloud may be done in a number of ways. In the view of [5,] it proposes an architecture for optimising the distribution of resources in cloud systems based on infrastructure as a service. The performance of distributed data-intensive applications hosted on existing IAAS systems may be affected by a lack of understanding of the hosted application's needs. This resource allocation problem may be solved by implementing an architecture that employs a "what if" approach to guide the IaaS's allocation decisions. The concept utilises a prediction engine and a genetic algorithm to evaluate the performance of resource allocations and a large search space with a lightweight simulator. TARA reduced job completion times by as much as 59 % to application-independent allocation solutions. TARA's architecture is shown in Figure 2, which depicts the inputs used in the scoring process.

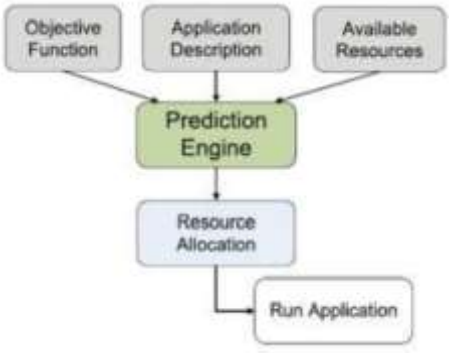

Figure: 2 Basic Architecture of TARA

b) Resource Allocation Resource Allocation Using a Linear Scheduling Strategy

It is feasible to provide the highest level of service to all customers [3] by employing the cloud environment's service node to handle all client requests and use all CPU, memory, and throughput resources. Waiting and response times increase when scheduling resources and tasks independently. Scheduling jobs and resources is handled by the LSTR scheduler, which uses a linear approach to the problem of resource allocation. KVM/Xen virtualization and LSTR scheduling are used in conjunction with one server node to create an IaaS cloud environment that maximises system performance and resource usage. There must be an integrated approach to resource consumption and allocation to maximise resource usage. The scheduling algorithms are primarily concerned with distributing the resources among the requestors in a way that maximises the given Quality of Service (QoS) characteristics. The cost function is the QoS parameter that we used in our assessment. The LSTR scheduling approach takes into account both the jobs and the number of virtual machines available. Because of this, resources will be used more efficiently.

c) Resources for Parallel Data Processing Allocated Dynamically

Efficient Use of Resources via Dynamic Allocation [6,7] presents a new cloud-optimized framework called Nephele for parallel data processing. Nephele is the first data processing framework which enable the dynamic allocation and de-allocation of various cloud computing resources during the execution of jobs and task scheduling. Virtual machines may be allocated to certain tasks in a processing job and automatically created and destroyed as the work progresses.

I. Benefits and Drawbacks of Resource Allocation Methods

Cloud computing offers various benefits in terms of resource allocation, regardless of the size of the firm or the business marketplaces in which it works. Due to the fact that this is a new technology, there will always be certain drawbacks. The benefits and drawbacks of allocating cloud-based resources will be discussed in the next paragraph.

Benefits:

- There is no need for a user to install any software or hardware in order to access, build, or host an application using a service like resource allocation.
- Another important advantage is that there are no restrictions on location or media. Our apps and data are accessible from any system, anywhere in the globe.
- There is no need for the user to invest in expensive hardware and software.
- A cloud provider's resources may be shared across the internet when there is a shortage of such resources.

Drawbacks:

- It is impossible for users to monitor the amount of resources they utilise since they rent them from remote servers.
- When a user wishes to migrate to a different service provider for better data storage, a migration issue arises. It's difficult to move large amounts of data from one service provider to another.
- Hacking and phishing assaults are possible in public cloud environments. It's simple for malware to proliferate since cloud servers are linked.
- Peripheral devices like printers may not be able to communicate with the cloud. Many of them need the installation of local software. Networked peripherals are less susceptible to errors and malfunctions.
- To properly allocate and manage cloud resources, a more comprehensive understanding of how the cloud works is necessary, and this expertise is mostly dependent on the cloud service provider.[8]

5. CONCLUSION

In this article, we've examined resource management as a whole, as well as contemporary research-based methodologies for allocating and monitoring resources. This paper summarises various theories and methods (algorithms) in order to have a better structure and model for allocation of resources and monitoring in order to enhance effectiveness, competitive nature, and efficiency in order to achieve the desired SLA, improved resource performance, and decreased power consumption. This study's main objective is to develop a new method for allocating and monitoring cloud computing resources.

REFERENCES

1. R. Buyya and R. Ranjan, “Special section: Federated resource management in grid and cloud computing systems,” Future Generation Computer Systems, vol. 26, no. 8, pp. 1189–1191, June, 2010.
2. K. Krauter, R. Buyya, and M. Maheswaran, “A taxonomy and survey of grid resource management systems for distributed computing,” Software: Practice and Experience, vol. 32, no. 2, pp. 135–164. 2002
3. V. Vinothina, R. Sridaran, and P. Ganapathi, “A survey on resource allocation strategies in cloud computing,” International Journal of Advanced Computer Science and Applications, vol. 3, no. 6, pp. 97– 104, 2012.

4. G. Aceto, A. Botta, W. De. Donato, and A. Pescapè, "Cloud monitoring: A survey," Computer Networks, vol. 57, no. 9, pp. 2093– 2115, 2013.
5. Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H. Katz, Topology aware resorce allocation for data-intensive workloads, ACM SIGCOMM Computer Communication Review, 41(1):120--124, 2011.
6. Daniel Warneke and Odej Kao, Exploiting dynamic resource allocation for efficient parallel data processing in the cloud, IEEE Transactions On Parallel And Distributed Systems, 2011.
7. Shenbaga Bharatha Priya, A., Ganesh, J. and Devi, M.M., 2014. Dynamic Load Rebalancing Algorithm for Private Cloud. In Applied Mechanics and Materials (Vol. 573, pp. 556-559). Trans Tech Publications Ltd.
8. Priya.A, Bhuvaneswaran R.S., 2021. Cloud service recommendation system based on clustering trust measures in multi-cloud environment. Journal of Ambient Intelligence and Humanized Computing, 12(7), pp.7029-7038.