Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Resource Allocation in Cloud Computing: A Review

Anju Bala
Computer Science Department, Shoolini University
India

Aman Kumar Sharma
Computer Science Department, HP University
India

Abstract— ***Cloud computing is a model that provides services through internet to its users. It enables users to put their data and services on the internet and same can access through internet. The most challenging task of cloud computing is resource allocation. In this paper, challenges of cloud computing are addressed in which resource allocation is one of the main challenges. Resource allocation and its challenges are discussed in detail.***

Keywords— ***Cloud Computing, Deployment models, Resource Management, Resource Allocation, Resource Allocation Challenges***

I. Introduction

Nowadays, computing is being introduced by a new paradigm called cloud computing. Cloud computing is an emerging paradigm in which computations are performed somewhere in a "cloud", which is a collection of data centres maintained and owned by a third party [1]. Cloud concepts ensures a cost effective realization of the utility computing principles, allowing users and provides easy access to resources in a self service, pay-as-you-go fashion thus decreasing cost for system administration and improving resource utilization [2]. Parallel processing, distributed processing and grid computing together emerged as cloud computing. It provides better services and data sharing through internet than distributed and grid computing. There are lots of definitions that define what is cloud computing? But the most acceptable is the definition that is given by National Institute of Science and Technology (NIST) and that is as follows, "... a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [3]." Cloud computing has three main service models provided to users known as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service) [4]. These are also known as layers of cloud computing.

A. Infrastructure as a Service

In cloud computing, on demand resources are virtually provided to consumers. The resources may be in the form of storage space, communication medium or computational capability [5]. In IaaS users are free to perform various activities to the server. Infrastructure services are considered to be bottom layer of cloud computing system [6].

B. Platform as a Service

In addition to IaaS, another approach is there to provide a higher level of abstraction to make a cloud easily programmable, that is known as Platform as a Service. It provides a platform on which developers create and deploy new applications and do not need to know how much memory and how many processors that applications will be using. These new applications are also inhibits multiple programming models and specialized services [7].

C. Software as a service

Software as a service is considered to be the top layer of cloud stack. Applications are reside on this layer which can be accessed by end users through Web portals. So, large numbers of end users are shifted from locally installed computer programs to on-line software services that offer the same functionality [8].

In cloud computing to provide quality of services the four parameters should be satisfied those are efficiency, scalability, robustness and security. In order to ensure the achievement of these quality constraints various technological and environmental issues have to be addressed. These issues are internal as well as external. However the key concerns of cloud computing is as follows:

A. Resource Management

Resource management is a core function of a cloud system. Resource management is required to handle the largely growing data, users and resources so that administrator can easily cater for this diversity and scale. The mechanism of processing power distribution, or the amount of memory, operates in such a way that the system dynamically allocates these parameters according to customer requirements. It affects the three basic criteria for system evaluation: performance, functionality and cost. The strategies for cloud resource management associated with the three cloud delivery models, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), differ from one another [9].

B. Scalability

The scalability requirement arises due to the constant load fluctuations that are common in the context of web based services. Load variation occurs because of the unpredictable growth in usage of data and resources. So, there is a need of a scalable design that can easily cope with this fluctuating load. Scalability of a system is a fundamental challenge in the context of cloud computing [10].

C. Reliability

Major problem for cloud computing is to minimize failover to provide reliable services. Cloud uses data multi-transcript fault tolerant, the computation node isomorphism exchangeable and so on to ensure the high reliability of the service. Using cloud computing is more reliable than local computer [11].

D. Multi-tenancy

In a cloud environment, Services owned by multiple providers are co-located in a single data centre. The performance and management issues of these services are shared among service providers and the infrastructure provider which may lead to performance degradation [12]. The layered architecture of cloud computing provides a natural division of responsibilities: the owner of each layer only needs to focus on the specific objectives associated with this layer. However, multi-tenancy also introduces difficulties in understanding and managing the interactions among various stakeholders [13].

II. RESOURCE ALLOCATION

Cloud computing provides a way of deploying and accessing massively scalable resources on demand, in real time and at affordable cost. Resource allocation is most difficult and challenging task in cloud system. It is the process of assigning available resources to the needed cloud users and applications [14]. In cloud the resource allocation is based on the IaaS and takes place at two levels:

- When an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, to balance the computational load of multiple applications across physical computers
- When an application receives multiple incoming requests should be assigned to a specific application instance to balance the computational load across a set of instances of the same applications [14].

Resource allocation strategies are all about integrating cloud provider activities for allocating and utilizing resources within the limit of cloud environment so as to meet the needs of cloud application. The resource allocation strategies should satisfy the following criteria:

A. Resource Contention

Resource contention arises when two applications try to access the same resource at the same time.

B. Resource Fragmentation

Resource Fragmentation arises when the resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation.

C. Scarcity

Scarcity of resources arises when there are limited resources and the demand for resources is high. The multiple applications needed different types of resources such as cpu, I/O devices, memory and the techniques should satisfy that request.

D. Over Provisioning

Over provisioning of resources arises when the application gets surplus resources than the demanded ones.

To satisfy these criteria's input is required from both the users and providers [15].

Resource Allocation Parameters are High throughput, Maximum efficiency, QoS aware, SLA aware, Maximum Energy and Power consumption.

III. RESOURCE ALLOCATION CHALLENGES

The main purpose of resource allocation strategies is to supervise all the resource allocation mechanism responsible for attending to application requirements and for controlling cloud resources. The resource allocation challenges are divided into four categories.

A. Resource Modeling and Description

Resource modeling is a process used to define and analyze resource requirements needed to support the business processes within the scope of corresponding cloud systems in organizations. Therefore, the process of resource modelling involves professional resource modelers working closely with business stakeholders, as well as potential users of the cloud system. Resource modelling defines how the cloud functions and deals with infrastructural resources. There should be schemas in cloud systems to represent the virtual networks, virtual resources and virtual networks. In cloud system Resource description Framework (RDF) and Natural Description Language (NDL) are two main schemas [16]. RDF is a framework for representing information in the Web. The resource modelling and description incorporates a challenge in cloud system that is interoperability. The main goal of interoperability is to enable the seamless flow of data across clouds from different providers and also between applications inside the same cloud [17].

If a particular model is very detailed it means that its details are relevant and should not be disregarded, which in turn makes the optimization problem much more difficult to handle and solve. On the other hand, more details can give more flexibility and allow for a better usage of resources. This concept is called the granularity of the resources description.

B. Resource Offering and Treatment

After modelling the resources next step is to offer and handle them through an interface. Each interface should provide a way for developers which efficiently tell the requirements of their applications. The requirements should be represented in the form of service level agreement [18].

Handling resources requires implementation of solutions to control all the resources available in the cloud. Management and control solutions would offer a set of signalling protocols to set up switches, routers and hypervisors. To delegate these control tasks, each Cloud provider implements their own solution that descends, from data centre control solutions. New signalling protocols can be developed for integrated reservation of resources in Cloud.

The resource modelling is not entirely dependent on the way the resources are offered to developers [11].

C. Resource Discovery and Monitoring

Success of cloud system depends on the efficient use of right resources. Cloud system is so dynamic in nature, that resource discovery and monitoring is a crucial step in resource allocation. Resource discovery is the process of finding suitable resources to perform a task and monitoring is the process of observing resources or services to track their status and purposes [19]. Searching and locating resources for executing jobs in a reasonable time in spite of the dynamicity and large scale of the environment is a very time consuming process and can decrease the performance of a system. So, the resource discovery and monitoring should be continuous and should be done efficiently.

Resource discovery is implemented through a discovery framework. This framework helps brokers to discover and match available resources, and typically is composed of distributed repositories, which are responsible for storing information about and descriptions of physical and virtual resources.

Resource monitoring should be continuous so that process of on demand acquiring and releasing of resources can be properly done [16]. The monitoring process may be active or passive. It is active when the nodes are autonomous and may decide when to send state information to some central entity. It is considered to be passive when an entity that collects information from nodes continuously and has passive relation to all the nodes. Clouds make use of both alternatives simultaneously to improve the monitoring solution.

D. Resource Selection and Optimization

In cloud system there are number of resources are available for service based demand. To know about available resources and selecting the most suitable resource is an important aspect. Selection is the ability to scale the number of resources. The selection procedure depends upon the usage patterns of resources and service provider's preferences. The main goal of resource selection is to identify a list of resources which are available for end users. Hence resource selection procedure in the cloud should be scalable, where the desirable information is retrieved quickly even if there are a large number of resources involved [20].

The resource selection may be done using an optimization algorithm. There are number of optimization algorithms are available. Resource selection strategies are categorized in two categories that are a priori and a posterior according to the moment when the optimization techniques are applied. In priori techniques, the optimization strategy should aware to all the factors that affect the allocation process and thus first allocation solution is already an optimal solution. Whereas in case of posterior techniques, once an initial resource allocation is made, which can be suboptimal solution, the resource allocation strategy should manage its resources in a continuous way. Resource utilization and provisioning are dynamic in nature, so it is more interesting that the posterior optimization technique reach an optimal allocation first and are able to optimize the old request and readjust them according to new demand [16].

IV. CONCLUSIONS

Cloud computing is the paradigm that provides services through three main service models that are PaaS, IaaS and SaaS. The resource allocation is the service that is provided through IaaS platform and the most challenging task in cloud system. There are limited resources, locality restrictions, dynamic nature of resource demand, heterogeneity and environmental necessities, so an efficient resource allocation strategy is required that matches the cloud environment. Resource allocation in cloud aims to provide high performance. Resource allocation strategy should have high throughput, maximum efficiency, QoS aware, SLA aware, maximum energy and power consumption as output parameters to meet the needs.

Along with these parameters the challenges of resource allocation are also addressed to explore both performance and energy efficiency. Challenges in resource allocation are divided into four categories that are resource modelling and description, resource offering and treatment, resource discovery and monitoring and resource selection. The challenge is to know about the available resources, their offerings and selection of optimal resource that satisfies all the requirements of the request.

REFERENCES

[1] Nick Antonopoulos and Lee Gillam, Cloud Computing: Principles *System & Application*, Springer, 2010.

[2] L. Schubert and K. Jeffery, *Advances in Clouds*, Report of the Cloud Computing Expert Working Group. European Commission ,2012

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing, National Institute of Standards and Technology," *Information Technology Laboratory*, Technical Report Version 15, 2009.

[4] A.K. Sharma, A. Ganpati and A. Bala, "Cloud Computing: A data Security Framework," *International Journal of Computer Science Engineering and Information Technology, vol. 3, issue 5, pp- 19-26, 2013.*

[5] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *IEEE Internet Computing*, vol. 13, issue 5, pp-14-22, 2009.

[6] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The Eucalyptus open-source cloud-computing system," *in Proc of IEEE/*ACM (CCGrid 2009), pp. 124_131, 2009

[7] Appistry Inc., Cloud Platforms vs. Cloud Infrastructure, White Paper, 2009.

[8] L. Youseff, M. Butrico, and D. Da Silva, "Toward a unified ontology of cloud computing," *in Proceedings of the 2008 Grid Computing Environments Workshop*, 2008, pp- 1-10.

[9] Dan C. Marinescu, "Cloud Computing: Theory and Practice*", Elsevier Science & Technology Books* Newnes, 2013

[10] D. Agrawal, AEI Abbadi, S. Das and A. J. Almore, "Database scalability, elasticity and autonomy in the cloud," *database Systems for Advances Application,* Springer Berlin, 2011.

[11] (2009) Cloud computing development status [online]. Available: http://hi.baidu.com/mc625263041/ blog/ item/ 22aeb1d0a492bd309a50276b.html

[12] Chandrashekhar S. Pawar and R.B. Wagh, "A review of resource allocation policies in cloud computing," in Proc. NCETIT-2012, 2(3), pp-165-167, 2012.

[13] Q. Zhang, L. Cheng and R. Boutaba, "Cloud computing: state-of-the-art and research challenge," *Journal of internet services and applications*, vol. 1, Issue 1, pp-7-18, 2010.

[14] N. Krishnaveni and G. Sivakumar, " Suirvey on Dynamic Resource Allocation Strategy in Cloud Computing Environment," *International Journal of Computer Applications Technology and Research,* vol. 2, issue 6, pp-731-737, 2013

[15] V. Vinothina, R. Sridaran and P. Ganapathi, "A Survey on Resource Allocation Strategies in Cloud Computing," *International Journalof Advanced Computer Science and Applications,* vol. 3, issue 6, pp-97-104, 2012

[16] G. E. Goncalves, P.T.Endo, T. Damasceno, A. V. De A.P. Cordeiro, D. Sadok,J. Kelner, B. Melander and J. Mangs, " Resource Allocation in Clouds: Concepts, Tools and Research Challenges," *XXIX SBRC-Gramado-RS,* 2011.

[17] T. Dillon, C. Wu, and E. Chang, "Cloud Computing: Issues and Challenges", *In IEEE International Conference on Advanced Information Networking and Applications,* pp. 27-33, 2010.

[18] P. Patel, A. Ranabahu and A. Sheth,"Service Level Agreement in Cloud Computing". *In Cloud Workshop at OOPSLA,* 2009.

[19] M.D. Arcy, N. Miller, L. Pearlman, I. Foster and C. Esselman, "Monitoring and Discovery in a web Services Framework: Functionality and performance of the Globus toolkit's MDS4".

[20] C. Banerjee, A. Kundu, S. Bhaumik, R. Sinha and D. Rana,"Framework on Service based Resource Selection in Cloud Computing," *International Journal of Information Processing & Management,* vol 3, issue1 ,2012.

