RESEARCH ARTICLE OPEN ACCESS

Allocation Strategies of Virtual Resources in Cloud-Computing Networks

[1]K.Delhi Babu,[2]D.Giridhar Kumar
Department of Computer Science and Engineering,
SreeVidyanikethanEngg.College , Tirupati .

Abstract—
In distributed computing, Cloud computing facilitates pay per model as per user demand and requirement. Collection of virtual machines including both computational and storage resources will form the Cloud. In Cloud computing, the main objective is to provide efficient access to remote and geographically distributed resources. Cloud faces many challenges, one of them is scheduling/allocation problem. Scheduling refers to a set of policies to control the order of work to be performed by a computer system. A good scheduler adapts its allocation strategy according to the changing environment and the type of task. In this paper we will see FCFS, Round Robin scheduling in addition to Linear Integer Programming an approach of resource allocation.

Index Terms--Cloud Computing, Virtual Machine, Resource Allocation, Linear Integer Programming

I. INRTODUCTION

Cloud computing can be seen as an innovation in different ways. From a technological perspective it is an advancement of computing, which's history can be traced back to the construction of the calculating machine. While from a technical perspective, cloud computing seems to pose manageable challenges, it rather incorporates a number of challenges on a business level, both from an operational as well as from a strategic point of view.

One fundamental advantage of the cloud paradigm is computation outsourcing, where the computational power of cloud customers is no longer limited by their resource-constraint devices. By outsourcing the workloads into the cloud, customers could enjoy the literally unlimited computing resources in a pay-per-use manner without committing any large capital outlays in the purchase of hardware and software and/or the operational overhead there in. It enables customers with limited computational resources to outsource their large computation workloads to the cloud, and economically enjoy the massive computational power, bandwidth, storage, and even appropriate software that can be shared in a pay-per-use manner.

The main cloud computing attributes are pay per use, elastic self provisioning through software, simple scalable services, virtualized physical resources. Models, such as cloud computing based on Virtual technologies enables the user to access storage resources and charge according to the resources access. Cloud computing platforms are based on utility model that enhances the reliability, scalability, performance and need based configurability and all these capabilities are provided at relatively low costs as compared to the dedicated infrastructures. This new model of infrastructure sharing is being widely adopted by the industries. Industries experts predict that cloud Computing has right future in spite of changing technology that faces significant challenge.

Cloud computing is a complex distributed environment and it relies heavily on strong algorithms for allocating properly CPU, RAM and hard disk operations to end users and core processes in a mutual and shared system. Here comes the matter of resource accounting and there are two distinct alternatives. The first one is strictly usage-oriented where you have a limited number of units. Such units can be connected to CPU and/or Memory usage, time or they can be a compound indicator. This covers generally the idea of utility computing. As a whole it gives some flexibility but it is more expensive in the long term. The second alternative is capacity pre-allocation. In this case there are different plans with predefined constant resources dedicated CPU and Memory. This still gives flexibility to upgrade resources on demand but it also allows lower price for higher resource usage in the long term.

When talking about a cloud computing system, it's helpful to divide it into two sections: the front end and the back end. They connect to each other through a network, usually the Internet. The front end is the side the computer user, or client, sees. The back end is the "cloud" section of the system. The front end includes the client's computer (or computer network) and the application required to access the cloud computing system. Not all cloud computing systems have the same user interface.

In cloud computing environment, resource allocation or load balancing takes place at two levels. First, when an application is uploaded to the cloud, the load balancer assigns the requested process to physical computers, attempting to balance the computational load of multiple applications across physical computers. Second, when an application receives multiple incoming requests, these requests should be each assigned to a specific requested application instance to balance the computational load across a set of instances of the same requested application.

II. SIGNIFICANCE OF RESOURCE ALLOCATION

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module.

Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria as follows:

a) Resource contention situation arises when two applications try to access the same resource at the same time.
b) Scarcity of resources arises when there are limited resources.
c) Resource fragmentation situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]
d) Over-provisioning of resources arises when the application gets surplus resources than the demanded one.
e) Under-provisioning of resources occurs when the application is assigned with fewer numbers of resources than the demand.

Resource users' (cloud users) estimates of resource demands to complete a job before the estimated time may lead to an over-provisioning of resources. Resource providers' allocation of resources may lead to an under-provisioning of resources. From the cloud user's angle, the application requirement and Service Level Agreement (SLA) are major inputs to RAS. The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host applications [1] by RAS. The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources, it also poses a crucial problem in allocating and managing resources dynamically across the applications.

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments.

Cloud resources consist of physical and virtual resources. The physical resources are shared across multiple compute requests through virtualization and provisioning [1]. The request for virtualized resources is described through a set of parameters detailing the processing, memory and disk needs. Provisioning satisfies the request by mapping virtualized resources to physical ones. The hardware and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented.

The complexity of finding an optimum resource allocation is exponential in huge systems like big clusters, data centers or Grids. Since resource demand and supply can be dynamic and uncertain, various strategies for resource allocation are proposed. This paper puts forth various resource allocation strategies deployed in cloud environments.

A. Task Scheduling

User assigns the task to be executed over cloud. The task is received by cloud coordinator (CC). Cloud coordinator forwards the task over datacenters (DC). Datacenters contains number of unfixed hosts consisting of pool of virtual machines (VM). These hosts can be configured or deleted as per the demand.[2]

B. Resources Allocation

VMM asks for resources from the resource provider by sending the task requirements. Resource provider checks the availability of resources with Resource Owner.

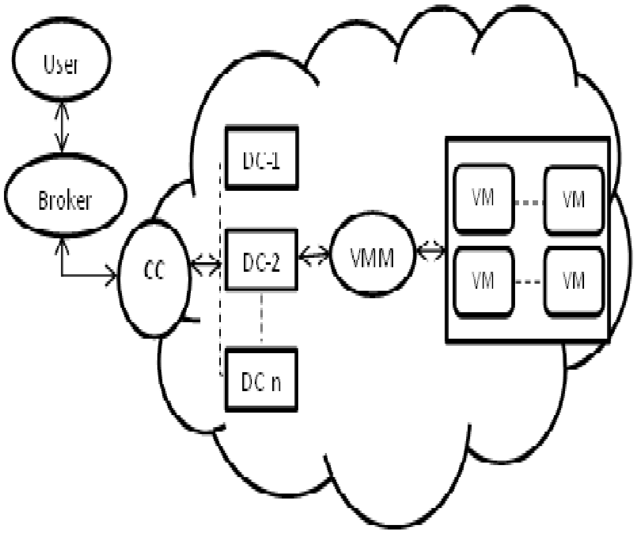

Fig.1 Task Execution

If the resources are available, the resource owner grants the access permission to use the resources to resource provider. Resource provider further provides access of the resources for creation of virtual machines.[5]

III. FIRST COME FIRST SERVE vs ROUND ROBIN

FCFS for parallel processing and is aiming at the resource with the smallest waiting queue time and is selected for the incoming task. Allocation of application-specific VMs to Hosts in a Cloud-based data center is the responsibility of the virtual machine provisioned component. The default policy implemented by the VM provisioned is a straightforward policy that allocates a VM to the Host in First-Come-First-Serve (FCFS) basis. The disadvantages of FCFS is that it is non preemptive. The shortest tasks which are at the back of the queue have to wait for the long task at the front to finish .Its turnaround and response is quite low. [6]

Round Robin (RR) algorithm focuses on the fairness. RR uses the ring as its queue to store jobs. Each job in a queue has the same execution time and it will be executed in turn. If a job can’t be completed during its turn, it will be stored back to the queue waiting for the next turn. The advantage of RR algorithm is that each job will be executed in turn and they don’t have to be waited for the previous one to get completed. But if the load is found to be heavy, RR will take a long time to complete all the jobs. The drawback of RR is that the largest job takes enough time for completion.

IV. LINEAR INTEGER PROGRAMMING

Linear programming is used for optimization problems and it is applied on specific problems with a particular formulation described in [7]. Best solutions like maximum gain or minimum cost is found through a mathematical model by the help of domain constraints encoded as linear equations. Linear programming methodology is widely used in operations research. Elements of linear programming are given below:

- Linear Objective Function: Value to be optimized is called objective function. This function should be represented as a linear equation, such as:

  Maximize: $c_1x_1 + c_2x_2$

- Constraints: Linear inequalities of the problem domain that bounds the solution space are called constraints. An optimum solution that meets these constraints’ requirements is searched by objective function. Each constraint should be represented as a linear equation, such as:

$$a_{1,1}x_1 + a_{1,2}x_2 \le b_1$$
$$a_{2,1}x_1 + a_{2,2}x_2 \le b_2$$
$$a_{3,1}x_1 + a_{3,2}x_2 \le b_3$$

Decision Variables: Both objective function and constraints are based on decision variables x_i whose optimal values are searched with simplex methods in linear programming.

Simplex Method: Basic algorithm generally used for linear programming is the simplex method [3]. It was proven to solve linear formulated problems of acceptable size in a reasonable time. The simplex method works by finding a feasible solution, and then moving from that point to any vertex of the feasible set that improves the cost function. Eventually a corner is reached from which any movement does not improve the cost function. This is the optimal solution. [3] The problem is usually formulated in matrix form, and represented as:

Maximize: $c^T x$
Subject to: $Ax \le b, x \ge 0$

where x represents the vector of variables (to be determined), c and b are vectors of (known) coefficients and A is a (known) matrix of coefficients [4]. In this formulation, a vector x is a feasible solution of the linear programming problem if it satisfies the given constraints. Problems defined in this formulation have three different types [3]:

- Infeasible: None of the vectors in solution space can satisfy the given constraints.

- Unbounded: Given constraints are not enough for bounding objective function parameters in the solution space. So, a better solution with an improved objective function value can exist.

- Optimal: Problem formulated in linear programming has an optimum value for the objective function and there exists vector(s) that can create such optimum value with satisfying the given constraints.

Integer linear programming is a customized version of linear programming where all decision variables of objective function are integers. As resource allocation problem requires integer results, only integer linear programming can be used for finding the solution.

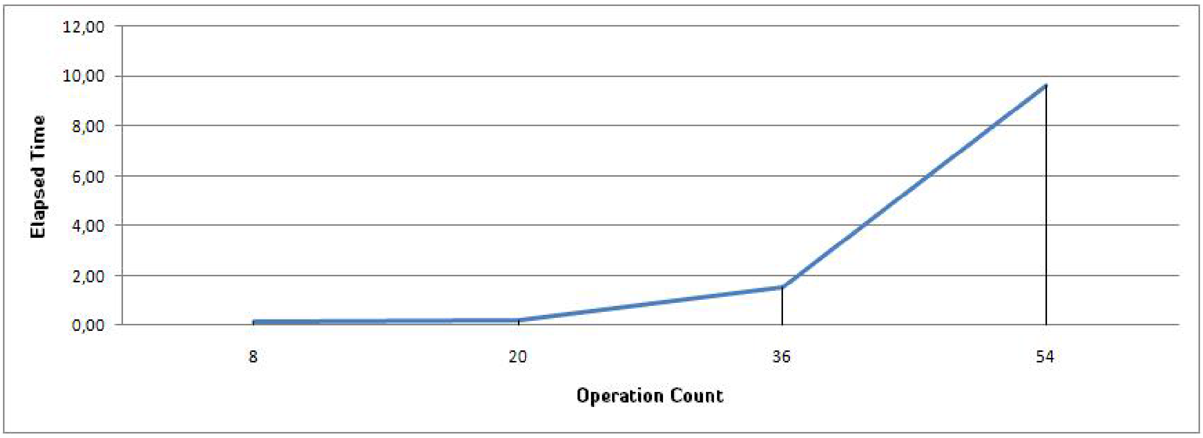

Fig 2: Elapsed Time for Integer Linear Programming Algorithm

V. ADVANTAGES AND LIMITATIONS

There are many benefits in resource allocation while using cloud computing irrespective of size of the organization and business markets. But there are some limitations as well, since it is an evolving technology. Let’s have a comparative look at the advantages and limitations of resource allocation in cloud.

A. Advantages:

1) The biggest benefit of resource allocation is that user neither has to install software nor hardware to access the applications, to develop the application and to host the application over the internet.
2) The next major benefit is that there is no limitation of place and medium. We can reach our applications and data anywhere in the world, on any system.
3) The user does not need to expend on hardware and software systems.
4) Cloud providers can share their resources over the internet during resource scarcity.

B. Limitations

1) Since users rent resources from remote servers for their purpose, they don’t have control over their resources.
2) Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It’s not easy to transfer huge data from one provider to the other.
3) In public cloud, the clients’ data can be susceptible to hacking or phishing attacks. Since the servers on cloud are interconnected, it is easy for malware to spread.
4) Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally. Networked peripherals have lesser problems.
5) More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

VI. CONCLUSIONS

Cloud computing technology is increasingly being used in enterprises and business markets. A review shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers.

This paper summarizes the main types of RAS and its impacts in cloud system. Some of the strategies discussed above mainly focus on memory resources but are lacking in other factors. Hence this survey paper will hopefully motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

REFERENCES

[1] Patricia Takako Endo et al. :Resource allocation for distributed cloud :Concept and Research challenges(IEEE,2011),pp.42-46 .

[2] 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.

[3] G. Dantzig, Linear Programming and Extensions. Princeton University Press, 1963. [Online]. Available: http://www.worldcat.org/isbn/0691059136

[4] N. Meggido, "Pathways to the optimal set in linear programming," Progress in Mathematical Programming Interior-point and related methods, pp. 131–158, 1988.

[5] Q. Cao, W. Gong and Z. Wei, "An Optimized Algorithm for Task Scheduling Based On Activity Based Costing in Cloud Computing," In Proceedings of Third International Conference on Bioinformatics and Biomedical Engineering, 2009, pp. 1-3

[6] V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, A survey on resource allocation strategies in cloud computing, International Journal of Advanced Computer Science and Applications, 3(6):97--104, 2012.

[7] C. Papagianni, A. Leivadeas, S. Papavassiliou, V. Maglaris, C. Cervello-Pastor, A. Monje, "On the Optimal Allocation of Virtual Resources in Cloud-Computing Networks," IEEE Teans. Computers, vol. 62, no. 6, pp. 1060-1071, Jun. 2013, doi: 10.1109/TC.2013.31.

K.Delhi Babu received the MS from BITS Pilani in Software Systems and PhD degree in Software Architecture from Sri Venkateswara University Tirupati, 2011. He is currently Working as Professor in Computer Science Department at Sree Vidyanikethan Engineering College, Tirupati , Andhra Pradesh. His current research interests include Software testing, Software Architecture and Software Engineering.

D.Giridhar Kumar received the B.Tech degree in Computer Science and Engineering, from Sree Vidyanikethan Engineering College affiliated to JNTUA, Tirupati, Andhra Pradesh, in 2011. He received M.Tech degree in Computer Science from the Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, 2014. His current research interests include network security and cloud computing.