IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-2, Issue-12 E-ISSN: 2347-2693

Study on Resource Allocation in Cloud

Swathy Surendran[1*], Sreetha E.S[2] and M. Azath[3]

*[1*2*3]Department of Computer Science and Engineering, Met's School of Engineering, Mala, India*

www.ijcaonline.org

Received: Nov /22 /2014 Revised: Nov/30/2014 Accepted: Dec/12/2014 Published: Dec/31/ 2014

Abstract—Cloud computing is an accepted trend in current computing which provide cheap and easy access to computational resources. In cloud computing, multiple cloud users can simultaneously request cloud services. Accessing of applications and associated data from anywhere is possible using clouds. Current cloud providers do not allocate the resources efficiently. Services are delivered to large number of users as demand grows up. This survey reviews various resource allocation methods**.**

Keywords- Cloud Computing, Computational Resources, Data Centers, Resource allocation, VM Provisioning and allocation

I. INTRODUCTION

Cloud focuses on maximizing the effectiveness of resource allocation. Cloud Computing refers to online manipulating, configuring, and retrieving the applications. It also offers online data storage, and data accessing. Cloud computing emerges as a new computing paradigm that provide reliable, customized and QOS (Quality of Service) for end-users[2]. Parallel processing, Distributed processing and grid computing together emerged as cloud computing. Cloud computing provide three types of services, including software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS). In Software as a service consumers obtain the capability to access and use a service that is hosted in the cloud. In Platform as a service users purchase access to the platforms, enabling them to make up their own software and applications in the cloud. In Infrastructure as a service Consumers control and manage the systems in terms of the applications, storage, operating systems, and connectivity of network, but do not control the cloud infrastructure. The main principle of cloud computing is that user data is not stored locally but is stored in the data center of internet. The cloud computing providers maintain and manage the operation of these data centers. At any time user can access the data by using Application Programming Interface (API) provided by cloud providers through any terminal equipment connected to the internet. Main goal of cloud computing is easy and scalable access of resources.

II. IMPORTANCE OF RESOURCE ALLOCATION

Providing the needed resources through the internet to the user is Resource Allocation method in cloud computing [3]. Proper allocation of resource is important. If the allocation is not managed accurately, then the functioning of entire cloud system will be collapsed. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. Resource Allocation Strategy (RAS) is integrating cloud provider activities to resources within the limit of meets the needs of the cloud application. To complete a user job, it requires the type and amount of resources needed by each application. We consider the resource allocation's order and time as the input for an optimal RAS. The criteria's that an optimal RAS should avoid are:

- Scarcity of resources.
- Over-provisioning
- Under-provisioning
- Scarcity of resources.
- Resource fragmentation

III . RESOURCE ALLOCATION METHODS IN CLOUD

A. Topology Aware Resource Allocation(TARA)

In cloud, a lot of resource allocation mechanisms are proposed. TARA [4] proposes an architecture for an optimized resource allocation in Infrastructure as a service (Iaas) based cloud system. Current Iaas systems significantly impact the performance for distributed data. TARA explains an architecture that uses a prediction engine and genetic algorithm. The performance of given resource allocation is estimated by this prediction engine with a lightweight simulator. Optimized solution in large search space is found out using this Genetic algorithm. Studies show that this is not completely an efficient mechanism.

B. Linear Scheduling Strategy for Resource Allocation

More waiting time and response time is required to Scheduling the resource and tasks separately. A named Linear Scheduling for Tasks and Resources (LSTR) scheduling algorithm[5] is designed to performs tasks and resources scheduling respectively. Here, a server node is used to establish the IaaS cloud environment and KVM/Xen virtualization along with LSTR scheduling to allocate resources which maximize the system throughput and resource utilization. Here Resource utilization is improved by integrating resource consumption and resource allocation. LSTR is designed to maximize the resource utilization. This is not so efficient since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

C. Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems

Resource management systems in data centers support Service Level Agreement (SLA)[6]-oriented resource allocation, SLA allocate resources according to the agreement. Based on the client's SLA, service provider allocates the resources. Here the method used is Force directed search algorithm. Total profit maximization is achieved but it does not ensure high efficiency and effective coordination of resource allocation.

D. Gossip

A gossip-based protocol is proposed for resource allocation in large-scale cloud environments [7]. For large clouds it performs a key function within distributed middleware architecture. Each machine in cloud environment is modeled as a dynamic set of nodes. Each of these nodes has a particular CPU capacity and memory capacity. The protocol implements a distributed scheme that allocates cloud resources to a set of applications that have time-dependent memory demands and it dynamically maximizes a global cloud utility function. When memory demand is smaller, the simulation results show that the protocol produces optimal allocation than the available memory in the cloud and the quality of the allocation does not change with the number of applications and the number of machines. But this work requires additional functionalities to make resource allocation scheme is robust to machine failure which spans several clusters and datacenters. Using this Gossip based method the organizations can cooperate to share the available resources to reduce the cost. Here the cloud environments of public and private clouds are considered. To obtain the optimal virtual machine allocation, they have formulated an optimization model. Related work is discussing that use desktop cloud for better usage of computing resources due to the increase in system utilization. The suggestion for a desktop cloud is that individual resource reallocation decisions using desktop consolidation and decision based on cumulative behavior of the system.

E. Automated Control of Multiple Virtualized Resources.

The method used is Auto Control: An automatic control [8] System. Main advantages are:

- Performance assurance
- All Applications can be meet their performance.
- Allocation decision should be made automatically without human intervention.
- Scalability can be achieved.
- Various workloads can be adopted.

Disadvantages of this method are:

- Auto Control only
- Does not deal the bottleneck problems.
- It does not control any memory

F. Adaptive energy-efficient scheduling.

The working of Adaptive energy-efficient scheduling(AEES) is based on AEES architecture[9] .The arrival of task in global queue is determine by the real-time controller and adaptive voltage controller and they determine if that task can be admitted or not. Scheduler assign a voltage level after accepting the task. For executing the admitted task, each node in the cluster maintains a local queue .The duty of local voltage controller is to minimizing the voltage levels for admitted tasks to reduce energy consumption. The scheduler follows following steps when a new task arrives:

1) System status information's such as: tasks running on the nodes, node voltage levels, waiting tasks in the local queues, execution times of finished tasks are checked by the scheduler .
2) Whether or not the new task can be allocated to a node and completed within its deadline by the energy-efficient global scheduling algorithm (EEGS) is decided by the scheduler. EEGS schedule a new task to a node with lowest voltage. The new task will be dropped to the rejected queue if its deadline is not guaranteed. Otherwise it task will be transferred to a destination node.
3) The status information's such as node voltage, sequence of the new task, execution time of tasks waiting in this node are passed on to this destination node. After a task in one node is executed, the local voltage adjuster relies on the local voltage adjusting algorithm or LVA to dynamically reduce the node voltage subject to the timing constraints of tasks waiting in the local queue. The dynamic voltage scaling approach can achieve high energy efficiency of the heterogeneous clusters.

   AEES integrates two algorithms EEGS and LVA. EEGS is implemented in the scheduler and LVA implemented in local adjuster. Although this algorithm improves the adaptively and schedulability of real-time heterogeneous clusters, they are not considered as an efficient allocation mechanism.

G. Utility Function

The objective functions such as minimizing cost function, cost performance function and meeting QoS objectives are optimized by many proposals to manage VMs in IaaS .

The objective function is defined as Utility property. Utility property is selected based on dealings of response time, targets met and profit, number of QoS etc. There are some works [10, 11] that dynamically allocate CPU resources in order to meet QoS objectives. They first allocate requests to applications that have highest priority. The author Dorian proposed Utility (profit) based resource allocation for VMs that uses live VM migration as a resource allocation mechanism. By changing VM utilities or node costs, this controls the cost-performance trade-off. This work is mainly concentrated on scaling CPU resources in IaaS. Live migration is the process of making running virtual machines or applications between different physical machines without disconnecting the client or application. Memory, storage and network connectivity of the virtual machines are transferred from the original host machine to the destination. Live migration is used as a resource provisioning mechanism by a few but all of them use policy based heuristic algorithm to live migrate VM and it is difficult in the presence of conflicting goals.

H. Policy based resource allocation in IaaS cloud[12]

Simple resource allocation policies like immediate and best effort are used by most of the Infrastructure as a Service (IaaS) clouds. The policy that allocates the resources only if available is treated as immediate allocation policy, otherwise the request is rejected. Best-effort policy also allocates the requested resources if available otherwise the request is placed in a FIFO queue. It is difficult for a cloud provider to satisfy all the requests at a time due to finite resources .These problems are addressed by a resource lease manager known as Haizea. Haizea introduce complex resource allocation policies. Haizea uses resource leases as resource allocation abstraction and implements these leases by allocating Virtual Machines (VMs). The kinds of resource allocation policies supported by Haizea are: best effort, immediate, advanced reservation and deadline sensitive. While analyzing the experiments results, it shows that resource utilization and acceptance of leases compared to the existing algorithm of Haizea maximizes but not fully considered as an efficient method.

TABLE 1: Comparison of resource allocation methods.

Existing System	Methods	Disadvantages
TARA	prediction engine and genetic algorithm	Topology Aware Resource Allocation reduce the job completion time only up to 59%
Linear Scheduling Strategy for resource allocation	(LSTR) scheduling algorithm	Not suitable for interactive real time applications
multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems	Force directed search algorithm	Maximize total profit but less efficient
Gossip	gossip-based protocol	Sharing of available resources
Automated control of Multiple Virtualized Resource.	Auto Control: An automatic control system	Auto Control does not deal the bottleneck problems.
Adaptive energy-efficient scheduling	Integrates two algorithms: EEGS and LVA	Occurrence of components which are responsible for high amount of power Dissipation.
Utility Function	Live migration: Migrating application into another system.	Sending of the VM's memory will consume the entire bandwidth. Only consider the live migration among the well connected data center.
Policy based resource allocation in IaaS cloud	Four policies used Immediate, best effort, advanced reservation and deadline sensitive	Though resource utilization increases ,efficiency is not achieved

IV. CONCLUSION

Cloud computing technology is more and more being used in business markets. An effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes the different resource allocation methods and its impacts in cloud environment. Each method has its own merits and demerits. Hence this survey paper will confidently motivate the researchers to arise with an efficient resource allocation algorithms and framework to strengthen the cloud computing standard. Different resource allocation problems are identified studied and solved the problems by designing mechanisms and algorithms for them. The efficiency of resource allocation is achieved only when user who values the item the most, gets it and balancing the load. Auction based mechanism balanced the load by efficiently allocate resources. The finest method to obtain efficiency is Auction based resource allocation.

REFERENCES

[1]http://en.wikipedia.org/wiki/Cloud_computing
[2]Lizhewang,JieTao,KunzeM.,Castellanos,A.C,Kramer,D., Karl,w, "High Performance Computing and Communications",IEEE International Conference HPCC,2008,pp.825-830.
[3] V.Vinothina, Sr.Lecturer, Dr.R.Sridaran, Dean, Dr.PadmavathiGanapathi"A survey on resource allocation strategies in cloud computing" International Journal of Advanced Computer Science andApplications, Vol. 3, No.6, 2012
[4] 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
[5] Abirami S.P. and Shalini Ramanathan, Linear scheduling strategy for resource allocation in cloud environment, International Journal on Cloud Computing: Services and Architecture(IJCCSA), 2(1):9--17, 2012
[6] HadiGoudarzi, MassoudPedram, Multi-dimensional SLAbased Resource Allocation for Multi-tier Cloud Computing Systems, in Proceedings of IEEE International Conference on Cloud Computing (CLOUD),Washington DC USA, 2011.
[7] RerngvitYanggratoke, FetahiWuhib and Rolf Stadler: Gossip-based resource allocation for green computing in Large Clouds: 7th International conference on network and service management, Paris, France, 24-28 October, 2011.
[8] PradeepPadala, Kai-Yuan Hou Kang G. Shin, Xiaoyun Zhu, Mustafa Uysal, Zhikui Wang, SharadSinghal, Arif Merchant "Automated Control of Multiple Virtualized Resources", The University of Michigan, Hewlett Packard Laboratories.
[9] Xiaomin zhua,chuan Hea,Kenli Li,Xiao Qin "Adaptive energy-efficient scheduling for real time tasks on DVS-enabled heterogeneous clusters", J.Parallel Distrib. Comput, SciVerse ScienceDirect, 2012 Elsevier Inc.
[10] D. Gmach, J.RoliaandL.cherkasova, Satisfying service level objectives in a self-managing resource pool. In Proc. Third IEEE international conference on self-adaptive and self organizing system.(SASO'09) pages 243-253.IEEE Press 2009
[11] X.Zhu et al. Integrated capacity and workload management for the next generation data center. In proc.5th international conference on Automatic computing(ICAC'08),pages 172-181,IEEE Press 2008.
[12]. Amit Nathani, Sanjay Chaudharya, GauravSomani, "Policy based resource allocation inIaaS cloud", Future Generation Computer Systems28 (2012)94–103 doi:10.1016/j.future.2011.05.016

AUTHORS PROFILE

Swathy Surendran has completed B.Tech in Computer Science and Engineering from Met's school of engineering, Thrissur, Kerala, in 2012. Presently pursuing M.Tech in CSE from Met's school of engineering, Thrissur, Kerala.
Sreetha E.S has completed M.tech from Mahendra college of engineering, Salem. Presently working as a lecturer in Met's school of engineering, Thrissur, kerala.
Dr. M. Azath is Head of Department of Computer Science and engineering, Met's School Of Engineering, Mala. He has received Ph.D. in Computer Science and Engineering from Anna University in 2011. He is a member in Editorial board of various international and national journals and also a member of the Computer society of India, Salem. His research interests include Networking, Wireless networks, Mobile Computing and Network Security.