A Survey on Optimal Resource Provisioning in Cloud

Mr. D. Seenivasan,
Assistant Professor,
Department of CSE,
K.S. Rangasamy College of Technology,
Tiruchengode, Namakkal (DT).

M. Gomathi,
M.E-I year
Department of CSE,
K.S. Rangasamy College of Technology,
Tiruchengode, Namakkal (DT).

***Abstract* - The pool of data centres is known as "cloud". Cloud computing is a paradigm which services are offered via internet in a pay-as-you going manner. The service is provided with data centres. All fields of computing were computation is required on the popularity of hybrid cloud; a use of both private and public clouds can be done effectively for the requirement of the user. Resource provisioning depends on specific job and parameters such as scalability, turnaround time, fault tolerance, load balancing and quality of service. Resource provisioning technique available for better utilization of resource and to provide high quality of service.**

The other challenges of resource allocation are meeting customer demands and application requirements. Cloud environment have many kinds of resource provisioning options and available in to reduce the total paying cost and better utilizing cloud resources.

Keywords – Cloud computing Environment, Resource management, Resource provisioning, Dynamic Resource Allocation, Resource Scheduling.

I. INTRODUCTION

Resource available and utilization are major role in any field like scientific, campus experiments and so on. A resource can be purchased from the requirement of the present and estimating a maximum of what it can be in near future [1]. The demand is not static nor progressive in a defined way, it is always in a wave form sometimes require more resources than we would have anticipated to be the maximum peak . It less than the average need hence determining the actual peak need of the resource such as website traffic in holiday will always be higher than the rest of the time [1].

Cloud computing is combined form of parallel processing and grid computing [2]. Application Programming Interface (API) provided by cloud providers by any terminal equipment connected to the internet. It provided storage services hardware and software services are available to the business markets. The basic cloud-service model, providersof IaaS (Infrastructure as a service) offer computers, physical or virtual machines and other resources (assume in a virtual-machine image-library, virtual local area networks ,file-based storage, blocks and firewalls, IP addresses, load balancers,and some of the software bundles[3]).

The services providers provide, and it can be everything, from the infrastructure, platform or software resources [2]. Cloud computing provided by Google, IBM Microsoft, Amazon, etc,[4].The developers deploying applications by a central server from computer hosted. So all applications can access a large network of computing resources [4].

IaaS providers flexible cloud solutions according to the hardware requirements of customers; and then customers can run operating systems and software applications on virtual machine (VMs)[5].

Cloud infrastructure provider's maximum of their profits by fulfilling their requirement of the consumers with minimum infrastructure and maximum resource utilization [6].

Resource Provisioning techniques are provided in Section II. In section III, Resource Allocation Strategies is presented. In section IV is shown Resource Allocation Mechanism. Section V presents Service-Level Agreement. Section VI shows the Challenge of Resource Provision in Cloud. And finally it is concluded in section VII.

II.RESOURCE PROVISIONING TECHNIQUES

To the increasing demand for computing resources, complexity and the size of today's data centers are growing rapidly. And also cloud computing infrastructures are becoming very popular [Fig.1]. An immediately,we want to ask one question is how the resources in a cloud computing infrastructure to be managed in a cost-effective manner[5].

An autonomic resource management could lead the efficient resource utilization and fast response in the presence changing workloads [5].

Genetic Algorithm

Cloud system consisting of a virtual cluster of physical nodes VM consolidation are strives to use a minimal number of nodes to accommodate all VMs in the system, gives an important thing in saving resource consumption. QoS is usually delivered by a VC as a single entity in VC Environment[5].So It have no reason why VMs' resource

capacity cannot be adjusted as long as the whole VC is still able to maintain the desired QoS(Quality Of Service)[5].

A Genetic Algorithm (GA) has been designed and implemented in this optimized system state, i.e., VM-to-node mapping and the resource capacity allocated to each VM, so the optimize resource consumptions in the Cloud. The increase in the arrival rates of the incoming requests may cause the current VMs in the VC cannot satisfy the desired QoS level, and it needs to be created with desired resource capacity [5].

The Genetic algorithm will be triggered for following situations, which are termed as resource fragmentation:

1) GA has spare resource capabilities in active nodes. An active node is a node in which the VMs are serving requests.

2) The spare resource capabilities in every node are less than the capacity requirements of the new VM.

3) The total spare resource capabilities across all used physical nodes are greater than the capacities required by the new VM. Typically, a GA to encode the evolving solutions, and its performs the crossover on the encoded solutions. Moreover, fitness functions to be defined to guide the evolution direction of the solutions.

Reconfiguration Algorithm

The Cloud system hosts multiple Virtual Clusters to server different types of incoming requests. A Genetic Algorithm is developed to compute the optimized system state and consolidate resources.

A Cloud reconfiguration algorithm is then developed to transfer the Cloud from the current state to the optimized one computed by the Genetic Algorithm. The Cloud system needs to reconfigure the Virtual Clusters through transiting the system state. During the transition [5], various VM operations will be performed, such as VM creation (CR), VM deletion , VM migrations as well as changing a VM's resource capacities.

Finally the clouds can transist from the current system to new system. The transition time represents overhead and should be minimized [5].

Cloud Dynamic Scaling Mechanism

Dynamic scalability becomes more attractive and practical because of the unlimited resource pool in cloud computing .cloud providers mostly offer cloud management to enable users to control their purchased computing infrastructure programmatically, but few of them directly offers a complete solution for automatic scalability activities [5]. The policies automatically scales up and scales down Virtual Machine instances by two aspects of a cloud application. These are performance and budget. Based on performance perspective, cloud auto-scaling mechanism enables cloud applications for finish all submitted jobs within the desired deadlines. From cost perspective [5], it reduces user cost by acquiring appropriate instance types which incurs less money and shuts down unnecessary instances when they approach full hour operation [5].

Integer programming is used to identify the most cost-effective instance types based on the job composition information of incoming workload, unexpected VM startup delay could not only affect the performance, it can also dominate the utilization rate, and therefore the cost, especially for short deadline cases.

Job processing time is also very important factors in our mechanism, because these two directly affect the number and type of provisioned instances [5].

III.RESOURCE ALLOCATION STRATEGIES

Allocation Strategies (RAS) proposed in cloud paradigm.

Execution Time

Resource allocation mechanisms are proposed in cloud. Actual task execution time is preemptable scheduling and its for resource allocation. Resource contention overcomes the problem and increases resource utilization by using different modes of renting computing capacities [2].

By estimating of execution time for a job is a hard task for a

user and errors are made very often[2]. And the VM model considered in is heterogeneous and proposed for IaaS.

Policy

The centralized user and resource management lacks in scalable management of users, resources and organization-level security policy [2], Dongwan et al. [2] has proposed a decentralized user and virtualized resource management for IaaS by adding a new layer.So refered as domain in between the user and the virtualized resources. Virtualized resources are allocated by role based access control (RBAC),to users through domain layer.

Virtual Machine (Vm)

A system can automatically scale and its infrastructure resources are designed in [2]. Virtual machines capable of live migration across multi- domain physical infrastructure [2].

IV. RESOURCE ALLOCATION MECHANISM

Now a day's many resource allocation strategies have come up in the literature of cloud computing environment and its technology has started maturing. Most number of Researcher communities around the world have proposed and Developed several types of resource allocation [4].

Topology Aware Resource Allocation reduced the job completion time of these applications by up to 59% when compared to application-independent allocation policies.

Topology Aware Resource Allocation (Tara)

The author mentioned in [4] the architecture for resource allocation in Infrastructure-as-a-Service (IaaS) based cloud systems. Current Infrastructure-as-a-Services of cloud providers are usually unaware of the hosted application's requirements and therefore allocate resources and its need independently ,which means,it can significantly impact performance for distributed data-intensive applications.

1) Architecture Of Tara

Topology Aware Resource Allocation [2] is composed of two major parts: a prediction engine and a fast genetic algorithm-based search technique. The prediction engine is the entity responsible for optimizing resource allocation. When prediction engine receives a resource request, the prediction engine works through the possible subsets of available resources (each distinct subset is known as a candidate) and identifies an allocation that optimizes estimated job completion time [4]. However, even with a lightweight prediction engine, iterating through all possible candidates is infeasible due to the scale of IaaS systems. Therefore a genetic algorithm-based search technique that allows TARA to guide the prediction engine through the search space intelligently is used.

2) Prediction Engine

The prediction engine maps resource allocation candidates to scores that measures their "fitness" with respect to a given objective function, TARA can compare and rank different candidates.

3) Objective Function

The objective function defines the metric that TARA should optimize. For example, In the data center, which is given the increasing cost and scarcity of power, an objective function might measure the increase in power usage due to a particular allocation.

4) Application Description

The application description have three parts: i) the framework type that identifies the framework model to use, ii) workload specific particular application's resource usage and iii) a request for resources including the number of VMs, storage, and so on.

5) Available Resources

The final input required by the prediction engine is a resource snapshot of the IaaS data centre. This includes information derived from both the virtualization layer and the IaaS monitoring service. The information gathered ranges from a list of available servers [4], current load and available capacity on individual servers to data centre topology and a recent measurement of available bandwidth on each network link.

Linear Scheduling Strategy In Resource Allocation

The processing time, resource utilization based on CPU usage, memory usage and throughput, the cloud environment with the service node to control all clients request.And It could provide maximum service to all clients [4]. Scheduling the resource and tasks separately involves more waiting time and response time.

A scheduling algorithm is Linear Scheduling for Tasks and Resources (LSTR) was designed, which performs tasks and resources scheduling respectively. In IaaS cloud environment, a server node is used to established and KVM/Xen virtualization along with LSTR scheduling to allocate resources which maximize the system throughput and resource utilization.

Parallel Data Processing Framework

Dynamic Resource Allocation is an Efficient Parallel data processing [4] introduces a new processing framework explicitly designed for cloud environments called Nephele. It is the first data processing framework to include the possibility of dynamically allocating/de-allocating different compute resources from a cloud in its scheduling and the job execution [4]. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution [4].

V. SERVICE-LEVEL AGREEMENT

A service-level agreement (SLA) is a service contract where various services are defined formally [6]. The term *SLA* is used to refer the contracted delivery time of the service or performance. Internet service providers will commonly include service level agreements within the terms of their contracts with customers to define the level(s) of service being sold in plain language terms [6].

SLA will typically have a technical definition in terms of mean time between failures (MTBF),mean time to repair or mean time to recovery (MTTR)[6] and then various data rates and throughput.

VI. CHALLENGE OF RESOURCE PROVISION IN CLOUD

Resource Provisioning has faces many challenges in cloud environment.

- Multiple cloud providers and service level agreement
- Multivariate uncertainty e.g., Price, Demands, Availability.
- Optimal solution under Uncertainty
- Computational Complexity

Provisioning Phases

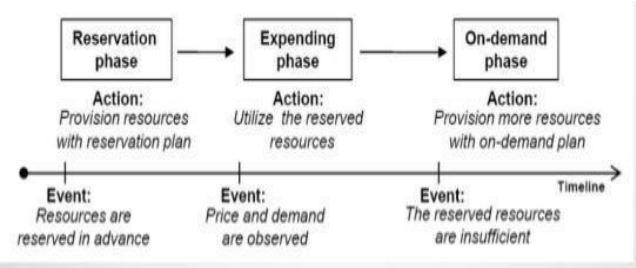

Fig.1

Reservation Contracts

Signed contract starting the time duration of availability of reserved resource [7]. Duration of contract period, price is to be discounted [Fig.2].

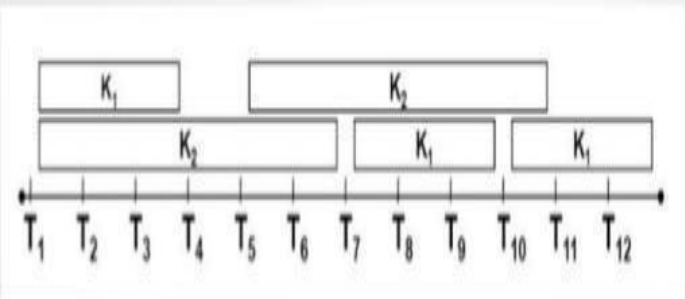

Fig.2

Optimal Solution under Uncertainty Uncertainty of Price

- On-Demand price might be fluctuated.

Uncertainty Oh Availability

- Free/cheap resources offered by cloud provider might be provided based on weak service level.
- Internet bandwidth is not reliable.

VII.CONCLUSION

Scheduling is one of the most important tasks in cloud computing environment. Cloud computing is being used in enterprises and business markets. It shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. In cloud paradigram [2], an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. The main type of RAS and its impacts in cloud system has used effiently.

These papers mainly focus on memory resources but are lacking in other factors. So this 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.

Cloud computing has been the hypothesis shift in distributed computing due to the way the resource provisioning and charging.Cloud Computing allow the users to efficiently and dynamically provision computing resource.Resource allocation is performed with the objective of minimizing the costs associated with it.

REFERENCES

[1] George Reese, "Cloud Application Architectures," Pub. O'Reilly Media, it-ebooks.info/book/286/, pp.1-10, 2009.

[2] A. Singh, M.Korupolu and D.Mohapatra. Server- storage virtualization: Integration and Load balancing in data centers. In Proc.2008 ACM/IEEE conference on supercomputing (SC'08) pages 1-12, IEEE Press 2008.

[3] G. juve and e. deelman, "resource provisioning options for large-scale scientific workflows," proc. ieee fourth int'l conf. e-science, 2008

[4] 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.

[5] Ambrust, A.Fox, R.Griffith, A.D.Joseph, R.Katz, A. Konwinski, G.Lee, D.Ratterson, A.Rabkin, I.Stoica and M.Zaharia,"A View of Cloud Computing," in communications of the ACM,vol.53,April 2010,pp.50-58.

[6] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, "Cloud computing andemerging IT platforms: vision,hype, and reality for delivering computing as the 5th utility",Future Generation Computer Systems 25 (2009) 599–616.

[7] Sharrukh Zaman, Daniel Grosu, **"**An Online Mechanism for Dynamic VM Provisioning and Allocation in Clouds", IEEE Fifth International Conference on Cloud Computing, 24-29 June 2012, pp 253-260, DOI 10.1109/CLOUD.2012.26.