RESEARCH ARTICLE OPEN ACCESS

A Survey on Resource Provisioning in Cloud

M.Uthaya Banu*, M.Subha**
*,**(Department of Computer Science and Engineering, Regional Centre of Anna University, Tirunelveli)

ABSTRACT
Cloud Computing allow the users to efficiently and dynamically provision computing resource to meet their IT needs. Companies are able to rent resources from cloud for storage and other computational purposes so that their infrastructure cost can be reduced. Further they can make use of company-wide access to applications based on pay-as-you-go model. Hence there is no need for getting licenses for individual products. However one of the major pitfalls in cloud computing is related to optimizing the resources being allocated. Resource allocation is performed with the objective of minimizing the costs associated with it. The other challenges in resource allocation are meeting customer demands and application requirements. In this paper we have presented a widespread survey on various resource allocation strategies and their challenges are discussed in detail.

Keywords – Cloud Computing, Infrastructure, Optimization, Resource Allocation, Virtualization

I. INTRODUCTION

Cloud Computing[1] refers to both the Cloud applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The datacenter hardware and software is what we will call a *Cloud*. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a *Public Cloud*; the service being sold is *Utility Computing*. We use the term *Private Cloud* to refer to internal datacenters of a business or other organization, not made available to the general public. According to NIST states “Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”. Three aspects are new in cloud computing

1)The illusion of infinite computing resources available on demand, thereby eliminating the need for cloud computing users to plan far ahead for provisioning.

2) The elimination of an up-front commitment by Cloud users, thereby allowing companies to start small and increase hardware resources only when there is an increase in their needs.

3) The ability to pay for use of computing resources on a short-term basis as needed (e.g., processors by the hour and storage by the day) and release them as needed, thereby rewarding conservation by letting machines and storage go when they are no longer useful.

Software, Platform, and Infrastructure[2] as a Service are the three main service delivery models for Cloud Computing. Those models are accessible as a service over the Internet. The Cloud services are made available as pay-as-you-go where users pay only for the resources they actually use for a specific time, unlike traditional services, Cloud Computing depends primarily on IaaS layer to provide cheap and pay-as-you-go processing power, data storage, and other shared resources.

The cloud computing model is comprised of a front end and a back end. These two elements are connected through a network. The front end is the vehicle by which the user interacts with the system and the back end is the cloud itself. The front end is composed of a client computer, or the computer network of an enterprise, and the applications used to access the cloud. The back end provides the applications, computers, servers, and data storage that creates the cloud of services. Cloud computing describes a type of outsourcing of computer services, similar to the way in which the supply of electricity is outsourced. Users can simply use it. They do not need to worry where the electricity is from, how it is produced, or transported. In cloud, services allowing users to easily access resources anywhere anytime. Users can pay for a service and access the resources made available during their subscriptions until the subscribed periods expire. Users are then forced to demand such resources if they want to access them also after the subscribed periods. IaaS providers build flexible cloud solutions according to the hardware requirements of customers; furthermore it let customers run operating systems and software applications on virtual machine (VMs).Customers merely pay for the resources that are actually used. To host web application services, service operators would apply resource subscription plans to dynamically adjust service capacity to satisfy a time-varying demand. While subscribing IaaS resources, the web service operators aimed to provide a certain level Agreement (SLA) with their clients, e.g., a guarantee on request response time. The resource provisioning of IaaS allows consumers to elastically increase or decrease the system capacity by changing configurations of computing resources. Moreover, cloud providers have multiple usage-based pricing model request from a cloud within a budgets based on different VM configurations, such as different CPU cores, memory size, and rental costs.

IaaS is the foundation layer of the Cloud Computing delivery model that consists of multiple components and technologies. Each component in Cloud infrastructure has its vulnerability which might impact the whole Cloud's computing security. Cloud Computing business grows rapidly despite security concerns, so collaborations between Cloud parties would assist in overcoming security challenges and promote secure Cloud Computing services.

II. RESOURCE PROVISIONING TECHNIQUES

To meet the increasing demand for computing resources, the size and complexity of today's data centers are growing rapidly. At the same time, cloud computing infrastructures are becoming popular. An immediate question is how the resources in a cloud computing infrastructure may be managed in a cost-effective manner. Static resource allocation based on peak demand is not cost-effective because of poor resource utilization during off-peak periods. In contrast, autonomic resource management could lead to efficient resource utilization and fast response in the presence changing workloads.

2.1 MicroEconomic-Inspired Approach

Infrastructure-as-a-Service[3] reduce the investment cost of renting a large data center while distributed processing frameworks are capable of efficiently harvesting the rented physical resources. Infrastructure as a Service (IaaS) have greatly reduced the investment risk of owning an infrastructure, but introduce a new performance scaling factor: the user's financial capacity to rent virtual resources. This additional factor complicates the deployment and management of cloud architectures. Combining resources from *IaaS*-Clouds with modern distributed computing frameworks allows for the effective handling of massively parallel problems. However, this combination introduces new challenges regarding both efficient use of cloud resources as well as user satisfaction. An answer to the key question of how many virtual machines (VMs) a user should request from an *IaaS* Cloud given that users have a limited budget and that there are speed-up barriers set by the available physical resources. Follow a microeconomic-inspired approach to determine the number of VMs allotted to each user according to user financial capacity. Since the underlying physical resources are shared among all cloud tenants, the performance the users get out of the cloud may significantly vary over time. Therefore, this approach continuously monitors the response time of user applications and adjusts the amount of resources accordingly. At its equilibrium point, the suggested approach maximizes profit. From the provider's point of view this profit corresponds to financial benefit whereas from the consumer's point of view, the same profit corresponds to quality of service received.

2.2 Genetic Algorithm

Virtual Clusters[4] are hosted in a Cloud system consisting of a cluster of physical nodes. VM consolidation, which strives to use a minimal number of nodes to accommodate all VMs in the system, plays an important role in saving resource consumption. In VC environments, QoS is usually delivered by a VC as a single entity. Therefore, there is no reason why VMs' resource capacity cannot be adjusted as long as the whole VC is still able to maintain the desired QoS. Treating VMs as being "mouldable" during consolidation may be able to further consolidate VMs into an even fewer number of nodes. A Genetic Algorithm (GA) has been designed and implemented in this work to compute the optimized system state, i.e., VM-to-node mapping and the resource capacity allocated to each VM, so as to 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 therefore a new VM needs to be created with desired resource capacity. The invocation of the GA will be triggered if the following situations occur, which are termed as resource fragmentation: 1) there are 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 then performs the crossover and the mutation operation on the encoded solutions. Moreover, a fitness function to be defined to guide the evolution direction of the solutions.

2.3 Reconfiguration Algorithm

The Cloud system[4] 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 by transiting the system state. During the transition, various VM operations will be performed, such as VM creation (CR), VM deletion (DL), VM migrations (MG) as well as changing a VM's resource capacities (CH).Finally the cloud will transist from the current system to new one. The transition time represents overhead and should be minimized. Experiments have been conducted to evaluate the reconfiguration algorithm.

2.4 Cost Aware Provisioning Approach

In[5] Cloud environments enable flexible, elastic provisioning by supporting a variety of hardware configurations and mechanisms to add or remove server capacity. This flexibility raises new challenges for application providers: (i) given several available resource configurations for a particular workload, which one to choose, and (ii) how best to transition from one resource configuration to another to handle changes in workload. a new approach for dynamically provisioning virtual server capacity that exploits pricing models and elasticity mechanisms to select resource configurations and transition strategies that optimize the incurred cost. Any dynamic provisioning algorithm involves two steps: (i) when to invoke the provisioning algorithm, and (ii) how to provision capacity so as to minimize infrastructure or transition cost.

2.5 Deterministic Resource Rental Planning

Optimization model [6] is based on a thorough rental cost analysis of running elastic applications in cloud. Considering the cost tradeoff between data generation and storage, developed a deterministic optimization model that minimizes the unit rental cost of covering customer demand. Based on this observation, we further designed a stochastic optimization model that seeks to minimize the expected resource rental cost given the presence of spot price uncertainty. Simulations based on realistic settings clearly demonstrate the advantages of our proposed optimization solutions in rental cost reduction.

2.6 Cloud Dynamic Scaling Mechanism

In [7] cloud computing, dynamic scalability becomes more attractive and practical because of the unlimited resource pool. Most cloud providers 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 in cloud. The mechanism automatically scales up and scales down VM instances by considering two aspects of a cloud application - performance and budget. From performance perspective, our cloud auto-scaling mechanism enables cloud applications to finish all submitted jobs within the desired deadline by acquiring enough VM instances. From cost perspective, it reduces user cost by acquiring appropriate instance types which incurs less money and shuts down unnecessary instances when they approach full hour operation. Integer programming is used to identify the most cost-effective instance types based on the job composition information of incoming workload, Long unexpected VM startup delay could not only affect the performance, but can also dominate the utilization rate, and therefore the cost, especially for short deadline cases. Workload and job processing time are also very important factors in our mechanism, because these two directly affect the number and type of provisioned instances.

2.7 Virtual Server Provisioning Algorithm

In [8] Amazon Elastic Compute Cloud (EC2) provides a cloud computing service by renting out computational resources to customers (i.e., cloud users). The customers can dynamically provision virtual servers (i.e., computing instances) in EC2, and then the customers are charged by Amazon on a pay-per-use basis. The challenge is how the customers efficiently purchase the provisioning options under uncertainty of price and demand. To address this issue, two virtual server provisioning algorithms are proposed to minimize the provisioning cost for long- and short-term planning. The objective is to minimize three cost namely expected costs, provisioning cost, variance cost and penalty cost. To compute the optimal number of reserved virtual servers can be achieved by formulating and solving the robust optimization model while the over-provisioning and under-provisioning problems can be avoided. For short-term planning, the optimal amount of bid server-hours i.e., optimal number of spot instances can be obtained by formulating and solving stochastic programming model. Also, the sample average approximation has been considered to address the complexity issue.

TABLE I
SUMMARY OF RESOURCE PROVSIONING TECHNIQUES

TITLE	TECHNIQUES USED	DESCRIPTON	LIMITATIONS	INFERENCE
Infrastructure as a Service Security: Challenges and Solutions.	Security Model for IaaS	Enhancing security in each layer of IaaS delivery model. SMI model consists:IaaS components, security model, and the restriction level.	i)IaaS layer to improve confidentiality and integrity of VMs ii) to get more controlled isolation environment	IaaS is the foundation layer of the cloud computing. Delivery model that consists of multiple components and presents an elaborated study of IaaS Components security and determines vulnerabilities and countermeasures.
Flexible Use of Cloud Resources through Profit Maximization and Price Discrimination.	Microeconomic Inspired Approach	It determines the number of VMs allotted to each user according to user financial capacity. This approach continuously monitors the response time of user applications and adjusts the amount of resources.	i)To reduce the time required to reach the maximum profit point. ii) to examine the effects of open- and closed-loop markets	Infrastructure-as-a-Service clouds reduce the investment cost of renting a large data center. This automatically adjusts to the ever changing equilibrium point caused by dynamic workloads and ensures that resources are shared proportionally.
Optimizing Resource Consumptions in Clouds	Genetic Algorithm	It has been designed and implemented to compute the optimized system state, i.e., VM-to-node mapping and the resource capacity allocated to each VM, so as to optimize resource consumptions.	i)It doesn’t provide optimized state when the node increases. ii)to guide the direction of the solutions a fitness function needs to be defined.	Developed a Genetic Algorithm to consolidate mould able VMs. In a virtualization based Cloud, two fundamental attributes of the system state are VM-to-node mapping and the resource capacity allocated to each VM. The developed GA performs the crossover and mutation operation on system states and is able to generate an optimized state. Moreover, the design of this GA is not limited to a particular type of resource, but is capable of consolidating multiple types of resource. A Simulator has been used to evaluate.

TITLE	TECHNIQUES USED	DESCRIPTION	LIMITATIONS	INFERENCE
Optimizing Resource Consumptions in Cloud	Cloud Reconfiguration Algorithm	It is developed to transfer the Cloud from the current state to the optimized one computed by the Genetic Algorithm .	i)Transition time at the high overhead. ii) the resource capacity allocated to the VMs in the node will exceed the node’s physical resource capacity after the reconfiguration.	Formalized a cost model to capture the overhead of a reconfiguration plan, and developed a reconfiguration algorithm to transit the Cloud to the optimized system state with the low overhead. A simulator has been developed to evaluate the transition
A Cost-aware Elasticity Provisioning System for the Cloud	Dynamic Provisioning algorithm	It involves two steps: (i) when to invoke the provisioning algorithm, and (ii) how to provision capacity so as to minimize infrastructure or transition cost.	i)To extend it with systems which employ queuing theory based model for capacity estimation for provisioning on cloud.	A approach for dynamically provisioning virtual server capacity that exploits pricing models and elasticity mechanisms to select resource configurations and transition strategies that optimize the incurred cost. Experimental Evaluation on public cloud using Amazon EC2.
Optimal Resource Rental Planning for Elastic Applications in Cloud Market	Deterministic Resource Rental Planning	An Optimization model is designed and it is based on a thorough rental cost analysis of running elastic applications in cloud.	i)Not suitable for the emerging spot instance. ii)To investigate optimum solutions for time varying workloads.	Designed an optimal resource rental planning model for elastic applications in a cloud environment. In particular, given known demand patterns over a specific time period, the ASP needs to periodically review the application progress so that no cost is wasted on excessive computation, data transfer and storage. Simulations over spot instance prices in Amazon EC2.
Cost Minimization for Provisioning Virtual Servers in Amazon Elastic Compute Cloud.	Virtual Server Provisioning Algorithm	To minimize the provisioning cost for long- and short-term planning. To compute the optimal number of reserved virtual servers and the optimal amount of bid server-hours.	Efficient strategies for bidding spot instances under price and demand fluctuation.	Proposed an virtual server provisioning algorithms to obtain the optimal solutions for purchasing the on demand, reservation, and spot options. Two algorithms are developed for long- and short-term planning.

III. CONCLUSION

Cloud computing has been the hypothesis shift in distributed computing due to the way the resource provisioning and charging. Managing Resource is a crucial task in making such an innovative technology to a larger consultation. Several researchers have put forward their ideas for new and innovative solutions for handling this imperative area. In this paper, we have carried out a decisive review of the most recent work carried out in this area.

REFERENCES

[1] 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.*

[2] W.Dawoud,I.Takouna, and C.Meinel, "Infrastructure as a Service Security: Challenges and Solutions," in *Proc the 7th International Conference on Informatics and Systems 2010(INFOS'10)*,Cairo,March 2010,pp.1-8

[3] K.Tsakalozos, H.Kllapi, E.Sitaridi, M.Roussopoulous, D.Paparas, and A.Delis,"Flexible Use of Cloud Resources through Profit Maximization and Price Discrimination," in *Proc of the 27th IEEE International Conference on Data Engineering(ICDE 2011)*,April 2011,pp.75-86.

[4] L.He,D.Zou,Z.Zhang,K.Yang,h.Jin and S.Jarvis,"Optimizing Resource Consumption in Clouds,"in *Proc. of the 12th IEEE/ACM International Conference on Grid Computing(Grid 2011)*,2011,pp.42-49.

[5] U.Sharma,P.J.Shenoy,S.Sahu, and A.Shaikh,"A Cost-Aware Elasticity Provisioning System for the Cloud",in *Proc.International Conference on Distributed Computing Systems*,July 2011,pp.559-570.

[6] S.Chaisiri,B.S.Lee and D.Niyato,"Optimization of Resource provisioning Cost in Cloud Computing,"*IEEE Transactions on Services Computing,Feb.2011,pp.164-177.*

[7] W-R.Lee,H-Y.Teng and R-H.Hwang,"Optimization of Cloud Resource Subscription Policy,"*IEEE Cloud Com 2012,Taipei,Taiwa, Dec 3-6,2012*

[8] Y.Hu,J.Wong,G.Iszlai,and M.Litoiu,," Resource Provisioning for Cloud Computing," in Proc. of the 2009 Conference of the center Studies on Collaborative Research,2009,pp.101-111.

[9] R.N.Calherios,R.Ranjan and R.Buyya,"Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments," in *Proc.International Conference on Parallel Processing*,Sep.2011,pp.259-304.

[10] H.Zhao,M.Pan,X.Liu,X.Li and Y.Fang,"Optimal Resource Rental Planning for Elastic Applications in Cloud Market," *the 26th IEEE International Parallel and Distributed Processing Symposium*, Shanghai, China, 2012.

[11] M.Mao,J.Liand M.Humphrey,"Cloud Auto-Scaling with Deadline and Budget Constarints,"in *Proc. of the ACM/IEEE International Conference on Grid Computing(GRID 2010)*,2010,pp.41-48.

[12] C.C.T.Mark,D.Niyato and C.Tham,"Evolutionary Optimal Virtual Machine Placement and Demand Forecaster for Cloud Computing,"in *Proc.International Symposium on Modelling,Analysis and Simulation Of Computer and Telecommunication Systems(MASCOT)*,July 2011,pp.85-95.

[13] S.Chaisiri,R.Kaewpaung,B.S.Lae and D.Niyato,"Cost Minimization for Provisioning Virtual Servers in Amazon Elastic Compute Cloud," in *Proc.IEEE International Conference on Advanced Information Networking and Applications(AINA)*,2011,PP.348-355

[14] S.Islam,J.Keung,K.lee and A.Liu,"An Empircal Study into Adaptive Resource Provisioning in the Cloud, "*Future Generation Computer Systems,vol.28,pp.155-162,Jan 2012.*

[15] E.Caron,F.Desprez and A.Muresan," Forecasting for Grid and Cloud Computing On-Demand Resources Based on Pattern Matching," in *Proc.Cloud Computing Technology ,2010,pp.456.*