Resource Management and Pricing Mechanism for Commercial Clouds

Rani.PR*1,Seenivasan.D*2
1*PG Scholar of Computer Science, K.S.Rangasamy College of Technology, Tiruchengode, India.*
Email: rranisenthil@gmail.com, Mobile No: +91 9600060447.
2*Assistant professor (Academic), K S Rangasamy College of Technology, Tiruchengode, India.*
Email:seenumoorthy@gmail.com

***Abstract*--Data sources and computational applications shares the resources from the cloud environment. Hardware, software and information are provided in cloud environment. Amazon and Elastic cloud computing (EC2) are the leading commercial cloud resource providers. Pricing schemes are used in commercial clouds. Resource provisioning contracts are designed with different time periods. The resource provider uses the Extended Optimal Cloud Resource Provisioning (E-OCRP) algorithm for resource optimization. The E-OCRP algorithm can provision computing resources for being used in multiple provisioning stages. The demand and price uncertainty is considered in E-OCRP. Deterministic equivalent formulation, sample-average approximation and Benders decomposition models are applied with OCRP. The dynamic price estimation and resource allocation process are supported by the enhanced E-OCRP model. The supply demand criteria are integrated with the E-OCRP. The pricing scheme is included with penalty and incentive models. Scenario reduction techniques are applied to reduce the number of scenarios.**

Keywords-- Cloud computing, resource provisioning, virtualization, broker architecture ,price management.

I. INTRODUCTION

Cloud computing is a large-scale distributed computing paradigm in which a pool of computing resources is available to users via the Internet. Computing resources, e.g., processing power, storage, software, and network bandwidth, are represented to cloud consumers as the accessible public utility services. Infrastructure- as-a-Service (IaaS) is a computational service model widely applied in the cloud computing paradigm. In this model, virtualization technologies can be used to provide resources to cloud consumers. The consumers can specify the required software stack, e.g., operating systems and applications; then package them all together into virtual machines (VMs). The hardware requirement of VMs can also be adjusted by the consumers. Finally, those VMs will be outsourced to host in computing environments operated by third-party sites owned by cloud providers. A cloud provider is responsible for guaranteeing the Quality of Services (QoS) for running the VMs. Since the computing resources are maintained by the provider, the total cost of ownership to the consumers can be reduced.

In this paper, minimizing both under provisioning and over provisioning problems under the demand and price uncertainty in cloud computing environments is our motivation to explore a resource provisioning strategy for cloud consumers [1]. In particular, an optimal cloud resource provisioning (E-OCRP) algorithm is proposed to minimize the total cost for provisioning resources in a certain time period. To make an optimal decision, the demand uncertainty from cloud consumer side and price uncertainty from cloud providers are taken into account to adjust the tradeoff between on-demand and oversubscribed costs. This optimal decision is obtained by formulating and solving a stochastic integer programming problem with multistage recourse. Benders decomposition and sample-average approximation are also discussed as the possible techniques to solve the E-OCRP algorithm. Extensive numerical studies and simulations are performed, and the results show that E-OCRP can minimize the total cost under uncertainty. The major contributions of this paper lie in the mathematical analysis which can be summarized as follows:

- The optimal cloud resource provisioning algorithm is proposed for the virtual machine management. The optimization formulation of stochastic integer programming is proposed to obtain the decision of the E-OCRP algorithm as such the total cost of resource provisioning in cloud computing environments is minimized. The formulation considers multiple provisioning stages with demand and price uncertainties.
- The solution methods based on Benders decomposition and sample-average

approximation algorithms are used to solve the optimization formulation in an efficient way.

- The performance evaluation is performed which can reveal the importance of optimal computing resource provisioning. The performance comparison among the E-OCRP algorithm and the other approaches is also presented.

II. RELATED WORK

The resource provisioning strategies in distributed systems were addressed in [10]. An architectural design of on-demand service for grid computing was proposed. A profile-based approach to capture expert's knowledge of scaling applications was proposed in which extra demanded resources can be more efficiently provisioned. The concept of resource slot was proposed. The objective is to address uncertainty of resources availability. In [3], a binary integer program to maximize revenues and utilization of resource providers was formulated. However, did not consider uncertainty of future consumer demands. An optimization framework for resource provisioning was developed. This framework considered multiple client QoS classes under uncertainty of workloads. The arrival pattern of workloads is estimated by using online forecasting techniques. In [6], heuristic method for service reservation was proposed. Prediction of demand was performed to define reservation prices. K-nearest-neighbors algorithm was applied to predict the demand of resources. In contrast, our work specifies that demands are given as probability distributions. In addition, the price difference between reservation and on-demand plans was not taken into account in all works in the literature.

As virtualization is a core technology of cloud computing, the problem of virtual machine placement (VM placement) becomes crucial. The broker-based architecture and algorithm for assigning VMs to physical servers were developed. In [7], a resource management consisting of resource provisioning and VM placement was proposed. In [8], techniques of VM placement and consolidation which leverage min-max and shares features provided by hypervisors were explored. In [9], a dynamic consolidation mechanism based on constraint programming was developed. This consolidation mechanism was originally designed for homogeneous clusters. However, heterogeneity which is common in a multiple cloud provider environment was ignored. Moreover, did not consider uncertainty of future demands and prices. A dynamic VM placement was proposed. However, the placement is heuristic-based which cannot guarantee the optimal solution.

The optimal virtual machine placement (OVMP) algorithm was proposed [5]. This OVMP algorithm can yield the optimal solution for both resource provisioning and VM placement in two provisioning stages. First, the problem is generalized into the multiple stage formulation. Second, the different approaches to obtain the solution of computing resource provisioning are considered. Finally, the performance evaluation is extended to consider various realistic scenarios.

III. SYSTEM MODEL AND ASSUMPTION

A.Cloud Computing Environment

As shown in Fig. 1, the system model of cloud computing environment consists of four main components, namely cloud consumer, virtual machine (VM) repository, cloud providers, and cloud broker. The cloud consumer has demand to execute jobs. Before the jobs are executed, computing resources has to be provisioned from cloud providers. To obtain such resources, the consumer firstly creates VMs integrated with software required by the jobs. The created VMs are stored in the VM repository. Then, the VMs can be hosted on cloud providers' infrastructures whose resources can be utilized by the VMs. In Fig. 1, the cloud broker is located in the cloud consumer's site and is responsible on behalf of the cloud consumer for provision resources for hosting the VMs. In addition, the broker can allocate the VMs originally stored in the VM repository to appropriate cloud providers. The broker implements the OCRP algorithm to make an optimal decision of resource provisioning.

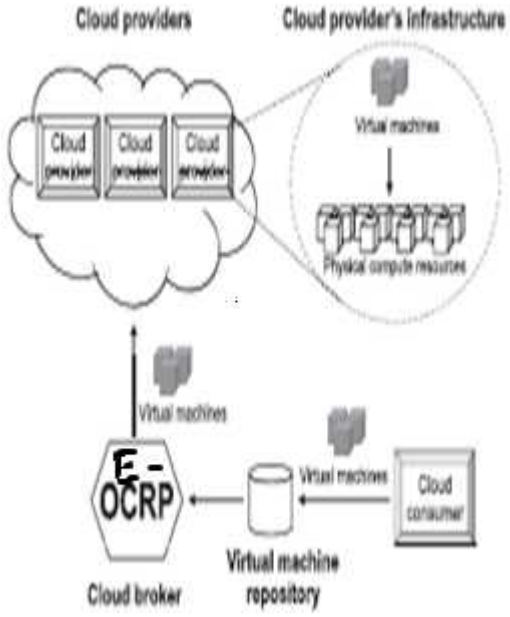

Fig. 1. System model of cloud computing environment.

In E-OCRP, there are multiple VM classes used to classify different types of VM. Let $I \subset IN_1$ denote the set of VM classes. It is assumed that one VM class represents a distinct type of jobs. A certain amount of resources is required for running the VM, and this required amount of resources can be different for VM in different classes. With this resource requirement, the cloud broker can reserve computing resources from cloud providers to be used in the future according to the actual demand. This demand can be determined as the number of created VMs. In this case, it is possible that additional resources can be provisioned instantly from cloud providers if the reserved resources is not enough to accommodate the actual demand.

Let J IN_1 denote the set of cloud providers. Each cloud provider supplies a pool of resources to the consumer. Let R denote the set of resource types which can be provided by cloud providers. Resource types can be computing power, storage and network bandwidth for Internet data transfer. Each VM class specifies the amount of resources in each resource type. Let b_{ir} be the amount of resource type r required by the VM in class $i \in I$. It is assumed that every cloud provider prepares facilities, e.g., virtualization management software, network facility, and load balancer, to support the consumer's hosting VM.

B. Provisioning Plans

A cloud provider can offer the consumer two provisioning plans, i.e., reservation and/or on-demand plans. For planning, the cloud broker considers the reservation plan as medium- to long-term planning, since the plan has to be subscribed in advance and the plan can significantly reduce the total provisioning cost [4]. In contrast, the broker considers the on-demand plan as short-term planning, since the on-demand plan can be purchased anytime for short period of time when the resources reserved by the reservation-plan are insufficient.

C.Provisioning Phases

The cloud broker considers both reservation and on-demand plans for provisioning resources. These resources are used in different time intervals, also called provisioning phases. There are three provisioning phases: reservation, expending, and on-demand phases. These phases with their actions perform in different points of time as follows. First in the reservation phase, without knowing the consumer's actual demand, the cloud broker provisions resources with reservation plan in advance. In the expending phase, the price and demand are realized, and the reserved resources can be utilized. As a result, the reserved resources could be observed to be either over-provisioned or under-provisioned. If the demand exceeds the amount of reserved resources, the broker can pay for additional resources with on-demand plan, and then the on-demand phase starts.

D.Provisioning Stages

A provisioning stage is the time epoch when the cloud broker makes a decision to provision resources by purchasing reservation and/or on-demand plans, and also allocates VMs to cloud providers for utilizing the provisioned resources. Therefore, each provisioning stage can consist of one or more provisioning phases. The number of provisioning stages is based on the number of planning epoches considered by the cloud broker, e.g., a yearly plan consists of 12 provisioning stages. Let $T \subset IN_1$ denote the set of all provisioning stages where $|T| \geq 2$. For resource provisioning under uncertainty, the broker is assumed to be able to reserve the resources in the first provisioning stage. Also, the broker obtains a solution, called recourse action, for provisioning resources against uncertainty parameters in every stage. These uncertainty parameters in each stage will be observed by the broker after the resource reservation has been made.

The observed uncertainty parameters are called realization. Then, the broker will take the recourse action according to the realization, e.g., utilizing the reserved resource and/or provisioning more resource with on-demand plan.

E. Reservation Contracts

A cloud provider can offer the consumer multiple reservation plans with different reservation contracts. Each reservation contract refers to the advance reservation of resources with the specific time duration of usage. For example, the reservation plan offered by Amazon EC2 has two reservation contracts [2], namely 1-year contract and 3- year contract. The certain amount of resources are reserved for 1 year in the one-year contract and 3 years in the three-year contract starting from the time when they are provisioned.

Let $K \subset IN_1$ denote the set of all reservation contracts which are offered by cloud providers. Let L_k denote the time duration specified in reservation contract $k \in K$. Let T_k denote the set of stages at which the cloud broker can provision resources by contract k. Let F_{kt} be the set of stages at which some resources reserved by contract k could be utilized at stage $t \in T$. Given the total number of stages |T|j, both T_k and F_{kt} are expressed as follows:

$$T_k = \{1,\ldots\ldots,|T|-L_k+1\}, \quad - \quad (1)$$

$$F_{kt}=\{\max(1,t-L_k+1),\min(t,|T|-L_k+1)\} \quad (2)$$

F.Provisioning Costs

With three aforementioned provisioning phases, there are three corresponding provisioning costs incurred in these phases, namely reservation, expending, and on-demand costs. The main objective of the OCRP algorithm is to minimize all of these costs while the consumer's demand is met, given the uncertainty of demand and price. For cloud provider, the price is defined in dollars ($) per resource unit. Let $c_{jkr}^{(R)}$ denote the unit price of resource type r subscribed to reservation contract k provided by cloud provider j in reservation phase of the first provisioning stage. It is assumed that the price of reservation plan in the first stage is charged by a fixed one-time fee.

IV. STOCHASTIC PROGRAMMING MODEL

The stochastic programming with multistage recourse is presented as the core formulation of the E-OCRP algorithm. First, the original form of stochastic integer programming formulation is derived. Then, the formulation is transformed into the deterministic equivalent formulation (DEF) which can be solved by traditional optimization solver software.

A.Stochastic Integer Programming for E-OCRP

subject to:

$$x_{ijk}^{(R)} \in \mathbf{N}_0, \quad \forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \forall k \in \mathcal{K}. \quad (3)$$

The general form of stochastic integer program of the E-OCRP algorithm is formulated in (3) . The objective function (3) is to minimize the cloud consumer's total provisioning cost. Decision variable $x_{ijk}^{(R)}$ denotes the number of VMs provisioned in the first provisioning stage. In other words, this number refers to as the total amount of reserved resources.

B. Deterministic Equivalent Formulation

Given a probability distribution of all scenarios in set Ω the formulation can be transformed into the deterministic integer programming called deterministic equivalent formulation. To solve this DEF, probability distributions of both price and demand must be available, i.e., p(ω). Then, the DEF can be solved by using traditional optimization solver software.

V. BENDERS DECOMPOSITION

The Benders decomposition algorithm is applied to solve the stochastic programming problem. The goal of this algorithm is to break down the optimization problem into multiple smaller problems which can be solved independently and parallelly. As a result, the time to obtain the solution of the E-OCRP algorithm can be reduced. The Benders decomposition algorithm can decompose integer programming problems with complicating variables into two major problems: master problem and subproblem.

VI. SAMPLE-AVERAGE APPROXIMATION

In the case that the number of scenarios is numerous, it may not be efficient to obtain the solution of the OCRP algorithm by solving the stochastic programming formulation directly if all scenarios in the problem are considered. To address this complexity issue, the sample-average approximation (SAA) approaches apply. This approach selects a set of scenarios, e.g., N scenarios, where N is smaller

than the total number of scenarios |Ω|. Then, these N scenarios can be solved in a deterministic equivalent formulation. The optimal solution can be obtained if N is large enough which can be verified numerically.

The SAA approach is applied to approximate the expected cost in every considered provisioning stage, i.e., $x_{ijkt}^{(R)}(\omega)$. A sampling method, is used to generate scenarios $\Omega_N = \{\omega_1, \ldots, \omega_N\}$, where N denotes the sample size. Let $N = \{1, \ldots, N\}$ be the set of indices of samples.

VII. RESOURCE AND PRICE MANAGEMENT FOR CLOUDS

The system is designed to manage the price and terms for resource sharing process. Prices and terms are dynamically estimated by the system. Supply/demand factors are considered in the pricing policy. The system is divided into five major modules. They are cloud providers, cloud consumers, resource provisioning, price optimization and term management.

Cloud providers module is designed to manage resources and providers. Cloud consumers module is designed to utilize resources under the providers. Resource provisioning module is designed to allocate resources based on user requirements. Price optimization module is designed to estimate resource price values. Term management module is designed to manage payment and resource usage terms.

A. Cloud Providers

The resources are provided by the cloud provider nodes. Different pricing methods are used for the resources. Resource requests are managed by the cloud broker. Resources are provided for the short term and long terms.

B. Cloud Consumers

The resource requests are submitted by the cloud consumers. Resources are requested with count and period information. Autonomous resource selection is done with price information. Resources are requested with demand and reservation plans.

C. Resource Provisioning

The resource provisioning is carried out with resource plans. Demand, expansion and reserve plans are used for the resource allocation. Resource cost is considered for the allocation process. The payments are collected in advance.

D. Price Optimization

The price optimizing is carried out to estimate dynamic prices for the resources. The supply/demand factors are considered in the price optimization process. Reservation prices are lower than demand prices. Local load and global loads are considered in the pricing policy.

E. Term Management

Term management is performed on long term jobs. The resource utilization period is divided into a set of terms. Job usage distribution is considered in the term estimation process. The term details are used for the payment process.

VIII. CONCLUSION

Cloud resources are provided on the basis of reservation and on-demand factors. Resource utilization cost for reservation plan is cheaper than on-demand plan. Cloud consumer can successfully minimize total cost of resource provisioning in cloud environments. The Extended Optimal Cloud Resource Provisioning (E-OCRP) algorithm is used to manage resource allocation and pricing process. E-E-OCRP algorithm can effectively save the total cost. Efficient term assignment mechanism is proposed in the system. Resource usage distribution is analyzed for term assignment process. Market based price assignment model is applied for price assignment under resource provider. The system minimizes the resource cost for users.

REFERENCES

[1] Sivadon Chaisiri, Bu-Sung Lee and Dusit Niyato, "Optimization of Resource Provisioning Cost in Cloud Computing" IEEE Transactions On Services Computing, Vol. 5, No. 2, April-June 2012.

[2] Amazon EC2, http://aws.amazon.com/ec2, 2012.

[3] A. Filali, A.S. Hafid, and M. Gendreau, "Adaptive Resources Provisioning for Grid Applications and Services," Proc. IEEE Int'l Conf. Comm., 2008.

[4] Amazon EC2 Reserved Instances, http://aws.amazon.com/ec2/reserved-instances, 2012.

[5] S. Chaisiri and D. Niyato, "Optimal Virtual Machine Placement across Multiple Cloud Providers," Proc. IEEE Asia- Pacific Services Computing Conf. (APSCC), 2009.

[6] K. Miyashita, K. Masuda, and F. Higashitani, "Coordinating Service Allocation through Flexible Reservation," IEEE Trans. Services Computing, vol. 1, no. 2, pp. 117-128, Apr.-June 2008.

[7] H.N. Van, F.D. Tran, and J.-M. Menaud, "SLA-Aware Virtual Resource Management for Cloud Infrastructures," Proc. IEEE Ninth Int'l Conf. Information Technology, 2009.

[8] M. Cardosa, M.R. Korupolu, and A. Singh, "Shares and Utilities Based Power Consolidation in Virtualized Server Environments," Proc. IFIP/IEEE 11th Int'l Conf. Symp. Integrated Network Management (IM '09), 2009.

[9] F. Hermenier, X. Lorca, and J.-M. Menaud, "Entropy: A Consolidation Manager for Clusters," Proc. ACM SIGPLAN/ SIGOPS Int'l Conf. Virtual Execution Environments (VEE '09), 2009.

[10] Y. Jie, Q. Jie, and L. Ying, "A Profile-Based Approach to Just-in- Time Scalability for Cloud Applications," Proc. IEEE Int'l Conf. Cloud Computing (CLOUD '09), 2009.