ABSTRACT

The usage of the cloud computing services is increasing rapidly across all the domains, thereby inducing the cloud service providers to increase their underlying cloud resource capacity. If service providers keep on increasing the capacity, it may not be an optimal utilization of resources. Since one of the main characteristics of cloud is optimal utilization of resources, we provide an analysis method, to aid both cloud service provider and cloud service consumer that both are benefitted from the cloud. Our first approach is to determine whether the resource request from user and actual usage is elasticity or new request may not be provisioned.

INTRODUCTION

Over the years, Cloud commuting is widely accepted as an effective service delivery model across all the industries. The conventional cloud architecture is customized to meet the requirement of the corresponding industries, which gave rise to different form of cloud such as community cloud, hybrid cloud, mobile cloud and so on. Various forms of cloud provide ways for creation, customization and delivery of services on cloud for multiple domains such as security as a service, authentication as a service, testing as a service, meta data as a service and so on. The evolution of various forms of cloud and variety of services delivered attracted more number of users from multiple domains on board into the cloud and the demand from users of multiple domains leads way to growth of variety of cloud services delivered over different forms of cloud. Both this scenario such as usage of cloud and multiple services on different forms of cloud leads to a situation for the cloud administrators to plan for his underlying capacity. Although public cloud environment provides better elasticity to the services when compared to the private cloud, the growth of the services through elasticity is limited to the capacity of the cloud. If the capacity is not planned properly by the cloud administrators, it may lead to the over deployment of underlying capacity or to the under deployment of the capacity. If the underlying cloud capacity is over deployed, it may lead to underutilization of resources on average. If under deployed, it may lead to degradation in quality of service as the demand for resources due to elasticity or new request may not be provisioned.

This leads to a situation of hybrid cloud where resources or services are fetched from other cloud service providers. The factors to be considered in this scenario of hybrid cloud are the service level agreement, security measures offered by the destination cloud services. At times, administrators should make a decision to include more resources into his cloud or to fetch resources and services from other cloud. Although the cloud users owns the ownership of the services that he avail from the cloud, he does not have control on the quality of service delivery, on scale up and scale down of resource utilization due to elasticity of his services, This makes variation in the cost of resource utilization thereby impacting user cloud budget. Our focus is to establish a solution to the above mentioned scenario with respect to the cloud administrators and users. Our solution provides a resource utilization prediction provides a mechanism to decide for the administrators to augment more resources into his cloud or to fetch resources from other cloud resources based on the resource utilization and demand pattern. This supply chain analysis assists the administrators to plan for his capacity thereby ensuring better quality and availability of resources with optimal utilization of resources. This decision making system considers the SLA, security and metering parameters of the target cloud resources. The capacity predictor for the users provides a predicted usage scheme, to plan for user cloud budget without compromising in the quality of service he avails from cloud. The outcome of the Decision making system will be aid the administrator to decide on any one of the following decisions such as the administrator to Switch on clusters from sleep mode or migrate services to other
cloud environment, provision request to public/private cloud environment, migrate existing VMs to other environment or to include more resources into the cloud environment. We also discuss the dynamic pricing scheme, which the administrator may offer to his users, thereby to increase the idle resource utilization and attract more users into the cloud.

Tian Guo et al., describes Seagull, a system designed to facilitate cloud bursting by determining which applications should be transitioned into the cloud and automating the movement process at the proper time. Seagull optimizes the bursting of applications using an optimization algorithm as well as a more efficient but approximate greedy heuristic, also optimizes the overhead of deploying applications into the cloud using an intelligent precopying mechanism that proactively replicates virtualized applications, lowering the bursting time from hours to minutes. Seagull’s placement algorithm considers both local reconsolidation opportunities and application cost characteristics, lowering the total cost of cloud bursting in response to data center overload by 45%. The approach focuses only on the application but not considered the platform and stack hosted in the cloud burst situation [1].

Cloud service providers face a challenge of measuring the service usage and charging the users. This problem is compounded as virtualization technology is deployed in many cloud platforms to consolidate servers and improve their utilization. The authors Jianzong Wang et al. provided three different but related models for apportioning costs in a private or public cloud environment supported by virtualized data centers. Through simulations and thorough comparisons of the results, authors finally champion the RO-BURST model tailored for the service providers’ need that is characterized by robustness and burstiness. Authors also import Cost Volatility Factors to ensure that the model is able to adjust itself to the market and multifarm demands in power and hardware components, such as disks and CPU, showing its compatibility and extensibility. The work lacks consideration of platform as a service [2].

Kikuchi, S et al., explored the characteristics of management operation performance by executing many operations simultaneously in an experimental virtual server system. The experimental results revealed some notable characteristics, including interference between operations with virtual machines (VMs) hosted on different physical servers and the asymmetric nature of live migration. The authors proposed a method to mitigate the management operation performance degradation or disturbance. In this method, VMs were manipulated according to proper scheduling policies that limit the operation concurrency [3].

Lu, Heng. et al., propose a workload forecasting strategy based on Gompertz curve to predict the sharp rising workload, and a customized resource management framework is also proposed to guarantee the high availability of the web applications and energy saving of the cloud service providers. The experiment first shows the accuracy of our workload forecasting model by using some workload statistics in the real world, and then a simulation-based experiment is designed to indicate that the proposed management framework detects changes in workload intensity that occur over time and allocates multiple virtualized IT resources accordingly to achieve high availability and energy saving targets. The techniques include a workload forecasting strategy based on Gompertz curve, a VM scheduler combined with VM provisioning, placement and resource collector. The target of our work is to help IaaS service providers to offer a better service for their customers, especially those whose applications may encounter a sudden rising workload [4].

Hong Xu and Baocun Li develop a general model that captures the resource needs of various applications and usage pricing of cloud computing. The approach considers alternative strategies that a provider can use to improve revenue, including resource throttling and performance guarantees, enabled by recent technical developments. The pricing analyses on several unique issues of cloud systems, including resource throttling, performance guarantees, and capacity right-sizing [5].

Rini T. Kaushik et al., presents a synergistic cloud where the pricing plan, scheduler, and the charge-back model work in tandem with each other to provide green and cost efficient computing options and incentives to the environmentally-friendly and cost-conscious end-users. The ne-grained, whole-sale electricity price aware scheduler works in synergy with end-users’ choices to reduce operating energy costs. The ne-grained, resource-utilization based, and spot price aware charge-back model is used to provide incentives in form of reduced usage costs. [6]. Few issues not addressed are considering job dependencies and cooling/thermal impact in the scheduler, discount models for providing incentives to users, and integrating of environmental costs in the charge-backs.

Linna Du. (2012) identified the problem of pricing driven virtual machine (VM) revenue maximization problem with Markovian traffics in a cloud market composed of hybrid cloud and public cloud. A pooled resource allocation model is built followed by numerical tests. The model could help cloud vendor decide which deployment model should be chosen. The cloud vendor could briefly allocate pooled resources on each virtual machine on each machine optimally. The dynamic resource allocation model could help cloud vendor to decide which type request she/he want to accept based on the data center utilization level. Also, the model could tell if the cloud vendor should expand the capacity to meet the demand. Price as a function of the quantity of pooled resources cloud implies the relationship of demand and price. This will help cloud vendor proposes better pricing strategy [7].

In order to help cloud providers achieving an agreeable price for their services and maximizing the benefits of both cloud providers and clients, Zhijie Li and Ming Li proposes a cloud pricing system consisting of hierarchical system, M/M/c queuing model and pricing model. Simulation results verify the efficiency of our
proposed system. A parameter $\alpha$ is set to control the workloads of each site. This approach can avoid the overload and reduce the pressures of whole system. At the same time, parallel system can significantly improve the execution efficiency [9].

Gueyoung Jung, et al., introduces a cloud recommendation platform, referred to as Cloud Advisor. It allows cloud users to explore various cloud configurations recommended based on user preferences such as budget, performance expectation, and energy saving for given workload. Then, it allows cloud users to compare offered price and performance with other clouds' offerings for the workload. By providing transparent comparisons, it can also support cloud provider to develop a competitive pricing strategy such as price reduction driven by energy efficiency. Using efficient constraints optimization algorithm, Cloud Advisor can address the tradeoff between these preferences and recommend near optimal cloud configurations. It has also integrated a benchmarking based approximation technique to characterize performances of target external cloud infrastructures for a given workload and then, allowed cloud users to compare offered price and performance with other clouds' offerings for the workload [10].

Sara Bouchenak survey existing work and identify gaps in existing cloud technology in terms of the verification tools provided to users. This author surveys the tools and methods that cloud users and service providers can employ to verify that cloud services behave as expected. We focus on the verification of several properties: the identity of the service and of the nodes the service runs on; functional correctness of a service; SLA-imposed parameters like performance and dependability; and lastly the compliance of the service with security requirements as specified by a security policy [11].

Terry Bleizeffer presents an overview of the user roles and basic framework. Then we will present a series of examples and best practices for interpreting and applying the roles in the design and operation of cloud computing solutions. The authors developed a framework for the definition of a single cloud user role as well as the taxonomy of the entirety of cloud roles. The framework established by the cloud user roles transcends these changes since roles are defined at a task-level and not at an individual job functional level. The relationships among business partners and resellers will be key influencers on the evolution of the cloud role definitions [12].

The authors aim to clearly portray the stringent and urgent need for applying privacy preserving approaches to the cloud and highlight the relevant work that has been done along these lines. The key objective is to identify important areas of user-cloud interaction and demonstrate a survey on the state of the art algorithms that have led to improved cloud privacy in these areas the focus is on exploring criteria for the impact of such approaches on user cloud interaction. Tanmay Sinha et al. intended to reiterate privacy preserving technologies in the light of different stages of user cloud interaction like access, storage, data mining, and personalization and then opened an intuitive discussion to analyze the merits and shortcomings of these technologies [13].

A sustainable production cloud environment along with its characteristics of elasticity ensures reliability and availability of services. To provide elasticity and high availability of existing services and to accommodate upcoming requests, the administrators should plan for the capacity and make his cloud ready to scale up automatically. The additional capacity for this auto scaling provision can be made in number of ways dynamically such as

- Increasing the underlying capacity of cloud.
- Fetch resources from other cloud.
- Migrate upcoming workload to other cloud.

Migrate existing workload to other cloud and accommodate new request in their own cloud.

To arrive at this decision, we need to make an estimate of the usage of resources. At first step we compare the user request for services with that of actual usage. For example if the user request for virtual machine in Infrastructure as a service, the request specification will be of {No. of VMs, RAM, Storage, No. of processor, OS}. The same will be provisioned to the users. An analysis is made at the requested configuration and actual usage of the provisioned resources through the following.

The trend analysis is generally carried out using aggregate data.

$$\ln(\text{ratei}) = \text{intercept} + (\text{slope1} \times \text{Yeari}) + (\text{slope2} \times \text{Var1i}) + (\text{slope} \times \text{Var2i})$$

(1)

In addition, statistical testing can be carried out to assess the current level of indicators with respect to the appropriate time objective.

$$z_1 = \frac{\text{Indicator}_{1\text{current}} - \text{Objective}_{1\text{current}} \times \text{multiplier}}{\sqrt{\text{n}_{1\text{current}}}}$$

$$z_2 = \frac{\text{Indicator}_{2\text{current}} - \text{Objective}_{2\text{current}} \times \text{multiplier}}{\sqrt{\text{n}_{2\text{current}}}}$$

(2)

Where $n_1$ and $n_2$ are the population denominators giving rise to indicator 1 and Indicator 2.

$$z_1 = \frac{\text{Indicator}_{1\text{projected}} - \text{Objective}_{1\text{projected}} \times \text{multiplier}}{\sqrt{\text{n}_{1\text{projected}}}}$$

$$z_2 = \frac{\text{Indicator}_{2\text{projected}} - \text{Objective}_{2\text{projected}} \times \text{multiplier}}{\sqrt{\text{n}_{2\text{projected}}}}$$

(3)

The analysis will provide the picture whether the resources are
under provisioned or
over provisioned or
optimally provisioned.

This will help the user to plan for their services accordingly to attain atmost quality of service.

The architecture of the cloud environment in interoperability scenario is provided with following Figure.

The architecture of the environment consists of the primary components such as Nodes to host the virtual machine services, Cluster head node for controlling clusters, Cloud head to manage the entire cloud environment, Security server to secure the entire cloud resources and services, Monitoring server that monitors the entire resources and services in cloud. The secondary components available are the capacity predictor to predict the capacity and pricing. The Metering server is used to vary the pricing scheme as dynamic pricing scheme. The interoperability server consists of Security server to take care of the security measures in interoperability scenario. Monitoring server monitors the resources and services at both the ends. The workflow server manages the flow of services at both ends.

In spite of all the mentioned above for additional capacity for scaling, the Infrastructure, platform services offered in the existing operational environment should have the primary clusters and secondary clusters for high availability. This requires the capacity of the resources hosting the services to be multiplied by double. Apart from this, for backup periodic backup irrespective of the backup mechanism, enormous storage space is required in the backup centre location. For operating services during disaster period, recovery resources are needed additionally. To meet the above mentioned situations, administrators deploy additional capacity to the cloud and keep them in spare. This additional capacity if avoided can compromise the quality of service hence it can’t be avoided. This additional capacity can bring additional overhead to the capital expenses, power supply, cooling factors, software licensing, operational expenses and other non IT infrastructure, network and so on. Since cloud computing is all about optimal utilization of the resources, this additional overhead is unfair. Hence our mechanism brings in an algorithm that provides the administrators a picture about the idle resources, whether to run the clusters inspite of the overheads or to make the clusters in sleep or hibernate mode. The decision to hibernate the clusters will be decided on certain parameters as mentioned below:

- the past resource utilization history on the whole history
- usage pattern with respect to time frame for elasticity
- prediction on upcoming requests

Other environmental / calendar factors

Based on the above mentioned criteria, our proposed resource prediction algorithm, will predict the capacity of the cloud resources required such as the servers, storage, network bandwidth, software licenses, platform and so on.

The following analysis procedure is employed to predict capacity planning; we employ the Smoothing with a finite moving average filter

\[ W_t = (2q + 1)^{-1} \sum_{j=-q}^{q} m_{t-j} + (2q + 1)^{-1} \sum_{j=-q}^{q} Y_{t-j} \approx m_t, \]

assuming that \( m_t \) is approximately linear over the interval \([t-q, t+q]\) and that the average of the error terms over this interval is close to zero. The spectral density is provided by

\[ f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-|h\lambda|} \gamma(h), \quad -\pi \leq \lambda \leq \pi, \]

After the prediction of capacity with the time frame whether the resource may be required for a short duration or for a longer duration, the administrators are forced to take any one of following decisions to meet the forecasted need of the computational and storage needs.

**Switch on clusters from sleep mode**

The administrator has got a first choice to wake up his resources from sleep mode which was planned and deployed for taking up elastic load.

**Move request to public/private cloud environment**

The second choice for the administrator is to forward the request received for his cloud environment to the other cloud environment governed by Service level agreements, security measures, metering pattern.

**Migrate existing VMs to other environment to accommodate new load**

The existing VMs offering less critical services running in the cloud environment can be migrated to other cloud environment without any loss in the session. This migration will free up few resources in the cloud to accommodate new upcoming requests and to meet elastic loads.
Add more resources into the cloud from other cloud

The required resources, although placed in the other cloud service provider location can be included into the cloud environment by taking up remote access. This ensures the resources owned by other cloud are directly controlled by the primary cloud.

Augment more resources into cloud

If the predicted capacity requirement is for a longer time, the administrator has got a chance to include new resources such as servers, storage and network into his existing cloud environment permanently. Selecting any one of the choices is to be made on analysis based on various parameters as follows:

- **Selection of option 1**, switch on servers from sleep mode for new request, the operational expenses on running the cluster, the licensing cost of the software used, need for servers in abnormal situation or sudden demand, elasticity needs, bandwidth available to take up the load is to be considered.

- **Selection of option 2 and option 3**, moving the request to other cloud environment and migrating existing service or virtual machines to other cloud environment, requires careful examination of the parameters such as the security mechanism offered by the target cloud environment, service level agreement, pricing policy, Monitoring level of the services, backup mechanism, network functionalities, meeting the scalability needs by target cloud, quality of service offered are to be considered.

- **Selection of option 4**, addition of servers from other cloud requires examining of the server benchmarking parameters, extending the security and monitoring capabilities from the primary cloud to those resources, latency, service level assurance that can be provided from those remote servers, flexibility are to be considered.

- **Selection of option 5**, augment the cloud with additional server, storage and network resources requires examining of the capital investment, usage of the servers with respect to time factor, return on investment, software licensing on those servers.

Poisson distribution process is employed as follows, the recurrence relationship is employed

\[
P(X = r) = \frac{\lambda}{r} P(X = r - 1) \quad \text{for } r \geq 1
\]

Above all to select any one of the above mentioned options, the target cloud environment and the primary cloud environment should be implemented based on standards at every layer from the network layer, virtualization, middleware, security and management layer and the APIs exposed by both the environment should be open at the management and security level.

From the cloud service consumer point of view, the users request for the cloud service, and their request is processed and delivered over virtualized resources. Service consumers at most of the times will have their IT budget preapproved to consume cloud services. The system administration department from the consumers should keep track of the service consumption, and preferable to know the predicted utilization and estimated billing to plan for their IT budget. Hence we propose the resource consumption schema for individual users based on the following criteria

- Input received from the users during the request for services,
- Service consumption pattern,
- Analysis on quantity and configuration of resources provisioned whether
- Resources are over provisioned
- Resources are under provisioned
- Criticality of the service request

The prediction of service requirement for individual users is obtained which can be utilized to calculate the predicted billing. To attract more service consumers into the cloud and to make return on investment for the cloud administrators on the resources by bringing in more service consumers, we introduce a varying pricing scheme for the cloud services, which is published by the administrators and consumers can avail the services on any of the schemes for their preference. The schemes are as follows:

- Fixed pricing scheme
- In fixed pricing scheme, the cost of the services is fixed for the consumers until the service contract is terminated.

Floating pricing scheme

In the floating pricing scheme, the cost of the services will be varying depending on the demand and supply arrived through supply chain management technique. The cost of the services and duration of the pricing scheme will be decided by the cloud service providers based on his analysis.

The pricing scheme should ensure that both the entities such as cloud service providers and the cloud service consumers are benefited. There are main factors that acts as enablers to determine the pricing scheme of cloud services apart from the demand and supply such as

- Hardware infrastructure which includes the investment for the servers, storage, network
- The non IT Infrastructure such as Power with backup power source, cooling facilities, the building infrastructure
- Manpower
- Operational expenses
- Software licenses
EXPERIMENTAL EVALUATION
The approach discussed was implemented in Cloudsim simulator, and the results are found to be supporting the approach. It is expected that the accuracy level of the prediction with respect to the actual on simulated environment. The cloud environment is established using the simulator by defining the servers, storage and network configuration. The request for cloud resources is provided to the simulator for the infrastructure and platform as a service. The pricing scheme is implemented and integrated into the simulator. The provision for dynamic pricing scheme is also integrated into the simulator.

CONCLUSIONS
Our proposed work to improve the availability of services, cater more users within limited resources, optimal utilization of resources, bring in the users to predict their cloud budget involves the follows process such as Continuous monitoring of the resource usage individually along with respect to elasticity, Prediction of request arrival pattern, Analysis on the requested resource usage pattern and analysis on the total resource usage pattern, based on these parameters to bring in optimal utilization, an option is provided to administrators whether to make the resources running or switch over to sleep mode.

Our proposed resource prediction algorithm, based on various criteria brings out the estimated requirement of additional servers, storage, network, software licenses and other tools. This prediction aids to decide on whether to switch on the clusters from hibernate mode, or to migrate the jobs to other cloud or to fetch resources from other cloud or to migrate existing workload to other cloud and accommodate new request in the same cloud.

FUTURE WORK
Our existing approach projects the prediction of cloud resources such as the IT resources in the cloud. Prediction of resource requirement can be extended to non IT resources such as electrical, cooling factors, manpower and infrastructure and so on. We have provided the basic prediction mechanism. As a future work, the prediction mechanism can be extended to an advanced level with inclusion of semantic analysis using ontology. The supply chain analysis can be further analyzed with suitable cloud customized algorithms.

REFERENCES


Trend analysis and Interpretation, key concepts and Methods for material and Child health professionals, HRSA, Maternal and Child health Bureau.