International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 3 Issue 1, January 2014

Study of cloud provisioning and it's Research challenges

Mitesh Barad#1, Prashant Jani#2, Richa Sinha#3

Abstract-**Efficient cloud provisioning allows the providers to effectively utilize their available resources and obtain higher profits. The trusted relationship between customer and provider is based on the guarantees to meet the SLA. Here we give you the idea about the three different provisioning techniques to handle the problems to fulfil the SLA requirement. A better solution would be to take into account user's demand when applying Cloud provisioning. Meeting the Quality of service (QoS) by customer is the key issue for the services provider to increase the profits. The perfect implementation of anyone cloud provisioning technique is the first step towards providing sufficient service to the user by maximum response time, low error rate, scalability, availability etc. Cloud provisioning is the solution to provide quality of services by the providers.**

***Index Terms*: SLA- service level agreement, QoS – Quality of service.**

I. INTRODUCTION

Cloud computing systems provide the next computing infrastructure enabling users to provision remote resources for their computational needs, eliminating the upfront costs of setting up their own systems. Clouds give users the illusion of an infinite computing resource available on demand and allow them to acquire and pay for resources on a short term basis. The usage model of cloud computing involves virtualization of computing resources. The cloud providers provision their resources into different types of virtual machine (VM) instances. These instances are then sold to the users for specific periods of time. Because cloud computing does not require any user level management and controlling on the low-level implementation of the system. Internet services can experience sudden, unexpected surge in demand, or the so-called flash crowd events. Those services are increasingly deployed in the cloud to take advantage of its auto scaling feature. A well-known example of the cloud model is the Amazon EC2 service which allows users to rent VM instances and operate them much like raw hardware. A key benefit of using such a service is the ability to provision a large number of VM instances when flash crowd happens.

This paper introducing the research challenges about cloud provisioning. The theory of different provisioning is explained. The static vision towards the cloud provisioning has been cleared by the paper explanation. The paper covered introduction about three types of cloud provisioning and research challenges of them.

II.CLOUD PROVISIONING

Cloud Provisioning [1] is the process of deployment and management of applications on Cloud infrastructures. It consists of three key steps [1]: (i) Virtual Machine Provisioning, which involves instantiation of one or more Virtual Machines (VMs) that match the specific hardware characteristics and software requirements of an application. Most Cloud providers offer a set of general-purpose VM classes with generic software and resource configurations. For example Amazon EC2 supports 11 types of VMs, each one with different options of processors, memory, and I/O performance; (ii) Resource Provisioning, which is the mapping and scheduling of VMs onto physical Cloud servers within a cloud. Currently, most IaaS providers do not provide any control over resource provisioning to application providers. In other words, mapping of VMs to physical servers is completely hidden from application providers; and (iii) Application Provisioning, which is the deployment of specialized applications (such as ERP system, BLAST experiments, and web servers) within VMs and mapping of end-user's requests to application instances.

A. VM PROVISIONING[1]:

The high level architecture of VM provisioning approach is shown in Figure 1. Different software components of the architecture are administered by the service provider. Its SaaS layer contains an admission control mechanism based on the number of requests on each application instance : if all virtualized application instances have k requests in their queues, new requests are rejected, because they are likely to violate negotiated maximum response time for an end user's request. Accepted requests are forwarded to the provider's PaaS layer, which implements the proposed system. Mainly, the following components are critical to the overall functionality of the system: (i) Application provisioner, main point of contact in the system that receives accepted requests and provisions virtual machines and application instances based on the

All Rights Reserved © 2014 IJARCET

input from workload analyser and from load predictor and performance modeller;

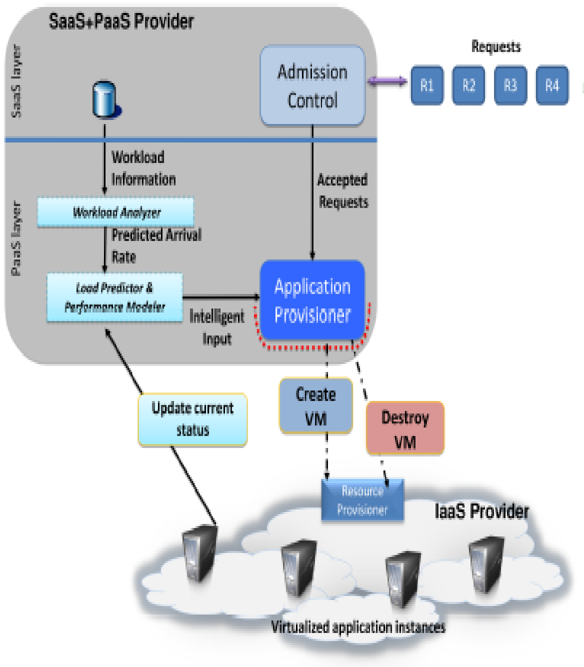

Figure-1. VM provisioning

(ii) Workload analyzer, which generates estimation of future demands for the application. This information is passed to the load predictor and performance modeler component; and (iii) Load predictor and performance modeler, which solves an analytical model based on the observed system performance and predicted load to decide the number of VM instances that should be allocated to an application.

B. RESOURCE PROVISIONING [2]:

Figure 2 shows the high-level architecture for supporting SLA oriented resource allocation in Cloud computing. There are basically four main entities involved:

- Users/Brokers: In general, the user interact with the Cloud management systems through an automatic systems such as brokers or schedulers who act on users behalf to submit service requests from anywhere in the world to the Clouds to be processed.

- SLA Resource Allocator: The SLA Resource Allocator acts as the interface between the Cloud computing infrastructure and external users/brokers[12]. It requires the interaction of the following mechanisms to support SLA-oriented resource management:

  - Service Request Examiner and Admission Control: The user service request is first interpreted by the Service Request Examiner and Admission Control mechanism that understands the QoS requirements before determining whether to accept or reject the request[12]. It ensures no SLA violation by reducing the chances of resource overloading whereby many service requests cannot be fulfilled successfully due to limited resources available. Therefore, it also needs the latest status information regarding resource availability (from VM Monitor mechanism) and workload processing (from Service Request Monitor mechanism) in order to make resource allocation decisions effectively. Then, it assigns requests to VMs and determines resource entitlements for allocated VMs.

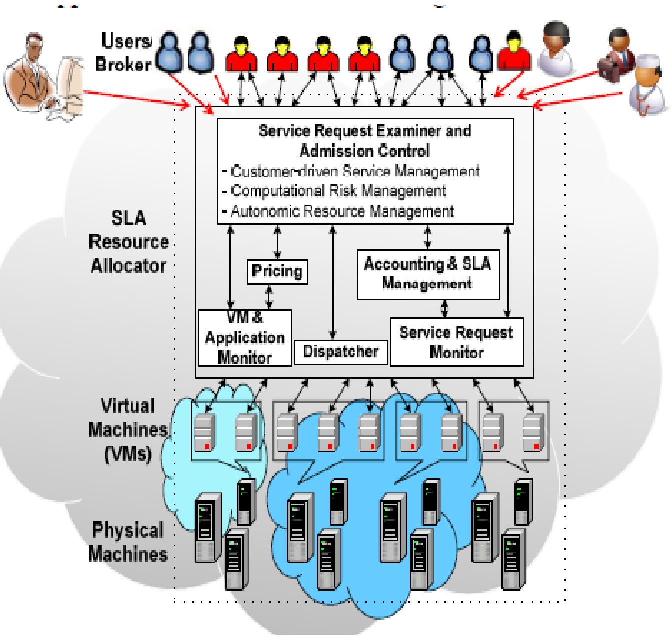
Users/ Broker
Service Request Examiner and Admission Control
- Customer-driven Service Management
- Computational Risk Management
- Autonomic Resource Management
SLA Resource Allocator
Pricing
Accounting & SLA Management
VM & Application Monitor
Dispatcher
Service Request Monitor
Virtual Machines (VMs)
Physical Machines

Figure-2. Resource provisioning

- Autonomic Resource Management: This is the key mechanism that ensures that Cloud providers can serve large amount of requests without violating SLA terms[12]. It dynamically manages the resources by using VM migration and consolidation. For instance, when an application requires low amount of resources, its VM is migrated to a host with lower capability, so that new requests can be served.

  - Pricing: The Pricing mechanism is a way to manage the service demand on the Cloud resources and maximize the profit of the Cloud provider. There are several ways in which service requests can be charged. For instance, requests can be charged based on submission time (peak/off-peak), pricing rates (fixed/changing) or availability of resources (supply/demand). Pricing also serves as a basis for managing computing resources within the data center and facilitates in prioritizing resource allocations effectively. Therefore, Cloud providers offer sometimes same/similar services at different pricing models and QoS levels. The two of the most prominent ones which are practically employed by Cloud providers: posted pricing and spot market.

- Accounting and SLA Management: SLA Management is the component that keeps track of SLAs of customers with Cloud providers and their fulfillment history. Based on SLA terms, the Accounting mechanism maintains the actual usage of resources by requests so that the final cost can be computed and charged from the users. In addition, the maintained historical usage information can be utilized by the Service Request Examiner and Admission Control mechanism to improve resource allocation decisions.

- VM and Application Monitor: Depending on the services provided, the resource management system has to keep the track of performance and status of resources at different levels. If service provided is compute resources, the VM Monitor mechanism keeps track of the availability of VMs and their resource entitlements. While in the case of application software services, the performance is continuously monitored to identify any breach in SLA and send a notification trigger to SLA Resource Allocator for taking appropriate action.

- Dispatcher: The Dispatcher deploys the application on appropriate virtual resource. It also takes the responsibility of creating Virtual machine image and their initiation on selected physical hosts.

- Service Request Monitor: The Service Request Monitor mechanism keeps track of the execution progress of service requests.

- Virtual Machines (VMs): Multiple VMs can be started and stopped dynamically to meet accepted service requests, hence providing maximum flexibility to configure various partitions of resources on the same physical machine to different specific requirements of service requests. In addition, multiple VMs can concurrently run applications based on different operating system environments on a single physical machine since every VM is completely isolated from one another on the same physical machine.
- Physical Machines: The data center comprises multiple computing servers that provide resources to meet service demands.

C. APPLICATION PROVISIONING[3]:

It is the deployment of specialized applications (such as ERP system, BLAST experiments, and web servers) within VMs and mapping of end-user's requests to application instances.

Application provisioning in cloud requires mechanism to automate and repeat as and when it is required. This is mainly because building blocks of an IT infrastructure such as computing, storage, network, security, is split into different security blocks and converted into a set of services. Now you can reach all of these services using a web service call rather than using GUI/Command line interface.

III.RESEARCH CHALLENGES

a. Customer-driven Service Management:

Yeo et al. [4] have highlighted customer satisfaction as a crucial success factor to excel in the service industry and thus proposed three user-centric objectives in the context of a computing service provider that can lead to customer satisfaction. However, there are many service quality factors that can influence customer satisfaction [5][6]. Factors that provide personalized attention to customers include enabling communication to keep customers informed and obtain feedback from them, increasing access and approachability to customers, and understanding specific needs of customers. Other factors that encourage trust and confidence in customers are security measures undertaken against risks and doubts, credibility of provider, and courtesy towards customers. Therefore, a detailed study of all possible customer characteristics needs to be done to determine if a data center needs to consider more relevant characteristics to better support SLA-oriented resource allocation.

b. Computational Risk Management:

Cloud computing is considered as the first fully accepted and implemented solution for providing computing as a utility. Having a commercial focused in offering computing services, there are several examples of elements in their resource management

[4] that can be perceived as risks. For example, if SLA with a customer is violated to fulfill Quality of a request of another customer, there is a risk of penalty and customer dissatisfaction. Hence, risk analysis from the field of economics can be identified as a probable solution to evaluate these risks. However, the entire risk management process [7][8] comprises many steps and thus need to be studied thoroughly so as to fully apply its effectiveness in managing risks. The risk management process comprises the following steps: establish the context, identify the risks involved, assess each of the identified risks, identify techniques to manage each risk, and finally create, implement, and review the risk management plan.

c. Autonomic Resource Management:

Service requirements of users can change over time and thus may require amendments of original service requests. As such, a data center must be able to self-manage the reservation process continuously by monitoring current service requests, amending future service requests, and adjusting schedules and prices for new and amended service requests accordingly. There are also other aspects of autonomy, such as self-configuring components to satisfy new service requirements. Hence, more autonomic and intelligent data centers are essential to effectively manage the limited supply of resources with dynamically changing service demand. For users, there can be brokering systems acting on their behalf to select the most suitable providers and negotiate with them to achieve the best service contracts. Thus, providers also require autonomic
resource management to selectively choose the appropriate requests to accept and execute depending on a number of operating factors, such as the expected availability and demand of services (both current and future), and existing service obligations.

d. SLA-oriented Resource Allocation Through Virtualization:

Recently, virtualization [9][10] has enabled the abstraction of computing resources such that a single physical machine is able to function as multiple logical VMs (Virtual Machines). A key benefit of VMs is the ability to host multiple operating system environments which are completely isolated from one another on the same physical machine. Another benefit is the capability to configure VMs to utilize different partitions of resources on the same physical machine. For example, on a physical machine, one VM can be allocated 10% of the processing power, while another VM can be allocated 20% of the processing power. Hence, VMs can be started and stopped dynamically to meet the changing demand of resources by users as opposed to limited resources on a physical machine. In particular, VMs may be assigned various resource management policies catering to different user needs and demands to better support the implementation of SLA-oriented resource allocation.

e. Service Benchmarking and Measurement:

Recently several Cloud providers have started offering different type of computing services. Therefore competition in the IT industry is increasing to maximize their market share. From the customer perspective, it is essential to have a service measurement standard to find out most suitable services which satisfy their needs. In this context recently, Cloud Service Measurement Index Consortium (CSMIC) has identified measurement indexes (Service Measurement Index - SMI) that are important for evaluation of a Cloud service [36]. For the performance evaluation of these services, there is essential requirement of real Cloud traces from various public archives such as PlanetLab and probability distributions to model application and service requirements respectively. This is because there are currently no service benchmarks available to evaluate utility-based resource management for Cloud computing in a standard manner. Moreover, there can be different emphasis of application requirements such as data-intensive and workflow applications, and service requirements such as reliability and trust/security. Therefore, it is necessary to derive a standard set of service benchmarks for the accurate evaluation of resource management policies. The benchmarks should be able to reflect realistic application and service requirements of users that can in turn facilitates the forecasting and prediction of future users' needs.

f. System Modeling and Repeatable Evaluation:

The proposed resource management strategies need to be thoroughly evaluated under various operating scenarios, such as various types of resources and customers with different service requirements in order to demonstrate their effectiveness. However, it is hard and almost impossible to perform performance evaluation of resource management strategies in a repeatable and controllable manner since resources are distributed and service requests originate from different customers at any time. Hence, we will use discrete-event simulation to evaluate the performance of resource management strategies. For our initial work [10], we have successfully used CloudSim [11] to evaluate the performance of resource management policies. CloudSim is a toolkit for modeling and simulation of Cloud resources and application scheduling. It allows easy modeling and simulation of virtual resources and network connectivity with different capabilities, configurations, and domains. It also supports primitives for application composition, information services for resource discovery, and interfaces for assigning application tasks to resources and managing their execution. Hence, these collective features in simulation toolkits such as CloudSim can be leveraged to easily construct simulation models to evaluate the performance of resource management strategies.

IV.CONCLUSION

In this paper, we pointed out many challenges in addressing the problem of enabling SLA-oriented resource allocation in data centers to satisfy competing applications demand for computing services. We envision the need for a deeper investigation in SLA oriented resource allocation strategies that encompass customer-driven service management, computational risk management, and autonomic management of Clouds in order to improve the system efficiency, minimize violation of SLA.

V.REFERENCES

[1] R. N. Calheiros, R. Ranjan, R. Buyya. Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments. 40th International Conference on Parallel Processing (ICPP) IEEE Computer Society, 2011, pp.295-304 DOI: 10.1109/ICPP.2011.17

[2] Rajkumar Buyya1,2, Saurabh Kumar Garg1, and Rodrigo N. Calheiros1 SLA-Oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions 2011 International Conference on Cloud and Service Computing

[3] CSS_Cloud_Enablement_Services 2010 CSS corporation.(s3.amazonaws.com/csslabs/docs)

[4] C. S. Yeo and R. Buyya. Integrated Risk Analysis for a Commercial Computing Service. Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium (IPDPS 2007), Long Beach, CA,USA, March 2007.

[5] B. Schneider and S. S. White. Service Quality: Research Perspectives.Sage Publications, Thousand Oaks, CA, USA, 2004.

[6] B. Van Looy, P. Gemmel, and R. Van Dierdonck, editors. Services Management: An Integrated Approach. Financial Times Prentice Hall,Harlow, England, second edition, 2003.

[7] M. Crouhy, D. Galai, and R. Mark. The Essentials of Risk Management. McGraw-Hill, New York, NY, USA, 2006.

[8] R. R. Moeller. COSO Enterprise Risk Management: Understanding the New Integrated ERM Framework. John Wiley and Sons, Hoboken, NJ, USA, 2007.

[9] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization.Proceedings of the 19th ACM Symposium on Operating SystemsPrinciples (SOSP 2003), pages 164–177, Bolton Landing, NY, USA,Oct. 2003.

[10] S. K. Garg, S. K. Gopalaiyengar, and R. Buyya. SLA-based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter. Proceedings of the 11th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP 2011),Melbourne, Australia, October, 2011.

[11] S. K. Garg and R. Buyya. NetworkCloudSim: Modelling Parallel Applications in Cloud Simulations. Proceedings of the 4th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2011),Melbourne, Australia, December, 2011.

[12] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Brogerg, and Ivona Brandic Cloud computing and Emerging IT platforms: Vision,Hype,and Reality for Delivering Computing as the 5th Utility