A Survey paper on Cloud Computing and its effective utilization with Virtualization

Anup H. Gade
Shri Ram Institute of Technology, Jabalpur
gadeanup@gmail.com

Abstract—Cloud computing delivers IT capabilities as services-on-demand. As the number of existing cloud vendors rises, resource count and types are ever increasing leading to a need of cloud management solutions which facilitate easy cloud adoption. While providing several services, cloud management's primary role is resource provisioning. In order to meet application needs in terms of resources, cloud developers must carefully choose among the existing offers in order to deploy their applications. This scalable and elastic model provides advantages like faster time-to-market, no capex and pay-per-use business model. While there are several such benefits, there are challenges in adopting public clouds because of dependency on infrastructure that is not completely controlled internally and rather shared with outsiders. Several enterprises, especially large ones that have already invested in their own infrastructure over the years are looking at setting up private clouds within their organizational boundaries to reap the benefits of cloud computing technologies leveraging such investments. Dynamic provisioning is a useful technique for handling the virtualized multi-tier applications in cloud environment. Understanding the performance of virtualized multi-tier applications is crucial for efficient cloud infrastructure management.

Index Terms— cloud management, cloud computing, resource brokering, resource provisioning; virtualized application

1 INTRODUCTION

Cloud computing [1] led to an innovative approach in the way in which IT infrastructures, applications, and services are designed, developed, and delivered. It fosters the vision of any IT asset as a utility, which can be consumed on a pay-per-use basis like water, power, and gas. This vision opens new opportunities that significantly change the relationship that enterprises, academia, and end-users have with software and technology. Cloud computing promotes an on-demand model for IT resource provisioning where a resource can be a virtual server, a service, or an application platform. Three major service offerings contribute to defining Cloud computing: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Infrastructure-as-a-Service providers deliver on-demand components for building IT infrastructure such as storage, bandwidth, and most commonly virtual servers, which can be further customized with the required software stack for hosting applications. Platform-as-a-Service providers deliver development and runtime environments for applications that are hosted on the Cloud. They allow abstraction of the physical aspects of a distributed system by providing a scalable middleware for the management of application execution and dynamic resource provisioning. Software-as-a-Service providers offer applications and services on-demand, which are accessible through the Web. SaaS applications are multi-tenant and are composed by the integration of different components available over the Internet. The offer of different models on which computing resources can be rented creates new perspectives on the way IT infrastructures are used, because Cloud offers the means for increasing IT resource availability whenever necessary, by the time these resources are required, reducing costs related to resource acquisition and maintenance. A case for exploring such a feature of Clouds is in Desktop Grids, which are platforms that use idle cycles from desktop machines to achieve high-throughput computing [2]. Typically, applications are executed in such platforms on a best-effort basis, as no guarantees can be given about the availability of individual machines that are part of the platform. If Desktop Grid resources are combined with Cloud
resources, a better level of confidence about resource availability can be given to users, and so it is possible to offer some QoS guarantees related to the execution time of applications at a small financial cost.

2 WHY CLOUD COMPUTING
Traditional infrastructure provisioning model is inefficient and does not meet the requirements of the internet era [Fig. 1]. In this system centric model, once the need for a business application is identified, its infrastructure needs are identified and a request for infrastructure is placed with the IT infrastructure team that procures and provisions the infrastructure. The application is then developed, tested and deployed on that infrastructure. Some of the challenges with this model include —

- **Need for Large Capex**: Large investments need to be made in procuring the infrastructure for a business application. This increases the barrier for innovation as it is hard to experiment with a business idea without large investments.
- **Poor Utilization of Resources**: Application usage is not going to be constant yet the infrastructure is provisioned for peak demand, to be able to guarantee application SLAs. So, the 12 infrastructure remains under-utilized for a major part of the time.
- **Slow Time-to-Market**: This model of procuring and provisioning infrastructure usually requires significant time and reduces the agility of an organization in creating new business solutions. Figure 2 below provides an overview of the service-centric provisioning model with cloud computing. In the cloud computing model, IT-related capabilities are made available as services that can be provisioned on demand. There are several offerings from various vendors that enable provisioning different IT components as services, components ranging from infrastructure to platforms and applications. This is commonly referred as infrastructure-as-a-service, platform-as-a-service and software-as-a-service. This cloud computing model offers several appealing benefits for enterprises including —
- **Faster Time-to-Market**: Enterprises can avoid the step of initial infrastructure procurement and setup, thus allowing the business solutions to be taken to market faster.
- **On-Demand Elastic Infrastructure**: Sudden spikes due to business growth, functionality additions or promotional offers can be addressed easily with Service Request Assign Request

3. MOTIVATION AND RELATED WORK
With the large scale adoption of cloud computing, where the essential characteristics are embraced and exploited by a larger pool of cloud providers and customers, the problem of resource allocation and management experienced a profound transformation from the traditional grid systems. In the case of the Grid Resource Management Systems (GRMS) the target was rather to obtain “high-throughput computation”, by reusing some idle resources, like in the case of Condor [6], or use some decentralized scheduling models, as in the case of Condor, or Legion [6], [7]. In the case of cloud computing, the process of negotiation and provisioning of resources is built around the principles of rapid elasticity and resource pooling, where “dynamic provisioning and reservation of computational resources” is one of the major concerns of different VM resource management solutions [8]. At the same time, on top of the measured service and on-demand self-service characteristics, strategies for market-based resource management systems are being reconsidered in the context of cloud computing [8], [9]. Different approaches exist for assuring scalability through negotiation and provisioning of cloud resources. Different SLA-based approaches for resource provisioning were considered. In [8], an SLA-oriented approach was considered for the Aneka, and CloudSim was used for performance evaluation. A policy-based approach for SLA-based negotiation was considered in [10], while [11] reconsider the problem of SLA-based provisioning by adding information about the response time, evaluated on Eucalyptus. Other SLA-based approaches were considered by [12], [13], or [14] in different application deployments.

A different approach, based on Quality of Service (QoS) maximization is offered in [15], [16]. The specific interest for scientific applications that was developed through grid systems, is exploited in conjunction with the cloud computing paradigm in various research papers. [17] describes an approach for elastic grid infrastructures, by employing a dynamic provisioning mechanism, while the approach from [18] is based on obtaining extra resources for highly resource-demanding scientific applications. On top of GrouSim, the work of [19] is oriented towards “analyze the problem of dynamic provisioning of Cloud resources to scientific workflows that do not benefit from sufficient Grid resources as required by their computational demands”. Approaches for on-demand resource provisioning are offered in [20], where specific...
time constraints are considered for essential activities. Elasticity and dynamic adaptation of services to user’s needs are considered in [21], while a similar approach, based on VM multiplexing is described in [22], [23] [24] aims to develop a fault tolerant environment, providing some “cost-aware and failure-aware provisioning policies”, with a significantly improved response time for user’s requests. Different platforms are used in the context of resource negotiation and provisioning, like Aneka and the CloudSim framework ([8], [25], [26], [18]), the OPTIMIS toolkit ([27], [28], [29]), or the Coasters system for automatically-deployed node provisioning ([30]). Different optimization approaches were considered in the context of resource provisioning. The Optimal Cloud Resource Provisioning algorithm was proposed in [31], as a stochastic programming model. A “feedback control based dynamic resource provisioning algorithm” is introduced in [15], considering a series of constraints, or QoS optimizations. The Automatic Resource Allocation Strategy based on Market Mechanism (ARAS-M) was specified in [32], where the mechanism is built around a QoS-based utility function, and a genetic algorithm is developed in close relation with this mechanism. While most cloud providers do not currently offer resource/service negotiation, according to Lomuscio et al: “automated negotiation will become the dominant mode of operation” [33]. Furthermore, by coupling automated negotiation with multi-agent systems we can make use of techniques from distributed systems and artificial intelligence. General agent-based approaches toward automated negotiation have been discussed in [34], [35].

4. AN EVOLUTIONARY APPROACH FOR RESOURCE PROVISIONING

Due to a combination of potentially large values, including the number of resource characteristics, associated policies and rules, the number of available cloud providers, as well as the number of exposed offerings, the classical approach for validation and testing of the negotiation model will be more computing intensive than an approach based on a similar genetic algorithm. In the evolutionary algorithm devised for building an SLA proposal, each gene represents an offer made by an identified vendor for a specific resource type, while the SLA proposal is a chromosome, made up of a set of genes. Two approaches have been considered in order to generate the initial population. In the first approach, the population is generated randomly using full chromosomes, as traditionally done in evolutionary algorithms. The second approach is a guided approach, where a chromosome contains only one gene from the gene pool and chromosomes are generated until all genes are covered. Standard genetic operators such as crossover, elitism or mutation are applied. The selection for crossover candidates is realized through a tournament selection. The method applied for crossover is uniform crossover. However, when using the guided approach, the crossover is slightly modified in order to determine the forming of chromosomes which have the maximum number of genes. This change takes place when one of the parents has a gene for a certain desired resource and the other one does not have one, thus always picking the gene over null, unlike traditional approach where there is a 50-50 chance between them.

A. Policies or rules

During the construction of a CfP, together with the list of desired resources different rules and policies can be specified, for guiding the negotiation process. The negotiation policies consist of a set of high level governing rules, specifying conditions/actions to be taken under specified conditions. One can use policies for filtering certain preferences for selecting best candidates under current SLA proposal. Thus, the role of policies is the allow the client greater flexibility in defining its preferences. Unlike other policies presented in the CfP, these are not restricting but mainly suggestive. Along with the importance assigned to each attribute of a resource, these policies help personalize the brokering outcome. The approach implemented in the prototype was to describe the rules as Jess2 rules and for each activated rule to add to a global variable the amount of fitness the rule gives. This, in turn, will be added to the overall fitness of a chromosome.

5. THE DYNAMIC VIRTUAL MACHINES IN CLOUD DATA CENTER

In order to dynamically provisioning resources for virtualized multi-tier application execution environments (VAEAs) of different customers, the most common approaches are based on self-managing techniques [9], such as Monitor, Analyze, Plan, and Execute (MAPE) control loops architecture is needed. The goal is to meet the virtualized application requirements while adapting IT architecture to workload variations. Usually, each request requires the execution of virtualized application allocated on the VM of each physical tier. A cloud data center enables multiple virtualized applications may be increased
when workload increases and reduced when workload reduces. This dynamic resource provision allows flexible response time in a VAEE where peak workload is much greater than the normal steady state. Figure 1 provides a high-level dynamic resource provision architecture for cloud data center, which shows relationships between computational resources pool and self-management community. Computational Resources Pool contains physical resources and virtualized resources. Plenty of VMs hold several VAEEs sharing the capacity of physical resources and can isolate multiple applications from the underlying hardware. VMs of each tier of a virtualized application may correspond to a physical machine. Computational resources pool delegates self-management community for satisfying the requirement goal of the customer to automatically allocate sufficient resources to the each tier of virtualized application. Self-management community means mechanisms to automate the VMs of configuring and tuning the virtualized multi-tier application so as to maintain the response time requirements of the different customers. It generates result of run-time provisioning for cloud data center. It includes four components as follows:

- **Monitor**: collects the workload and the performance metric of all running VAEEs, such as the request arrival rate, the average service time, and the CPU utilization, etc.
- **Analyzer**: receives and analyzes the measurements from the monitor to estimate the future workload. It also receives the response times of different customers.
- **Resource Scheduler**: sets up performance analytic models for each tier of the VAEE, and uses its optimizer with the optimization model to determine resource provisioning according to these workload estimates and response time constrains of different customer such that the resource requirements of the overall VAEE is minimized.
- **Virtualized application Executor**: assigns the VM configuration, and then runs the VAEEs to satisfy the resource requirements of the different customers according to the optimized decision. The goal is to minimize the using of resources under a workload while satisfying different customer for the constraints of average response time.

### 6. CONCLUSION & FUTURE WORK

The work presented in this paper addresses an important problem to cloud management solutions, that of providing adequate resources for cloud applications, during the negotiation process. The focus is on giving cloud developers the ability to express their preference towards resources, resource attributes and define relations between them. These relations are being defined through the use of policies it is argued that dynamic provisioning of virtualized multi-tier applications raises new challenges not addressed by prior work on provisioning technique for cloud environment. We presented an optimal autonomic virtual machine provisioning architecture for cloud data center. We proposed a novel dynamic provisioning technique, which was a hybrid model for a virtualized multi-tier application in cloud data center. A constrained non-linear optimization model is employed to minimize the total number of VMs for the requirement of customer. Hence the efficiency and flexibility for resource provisioning were improved in cloud environment. We evaluated and contrasted the performance of three tier virtualized applications through simulation experiments. Results have shown that under fine-grained resource provisioning, computing resources are optimized utilization. Moreover, our technique is also demonstrated that by optimizing provisioning the overall performance could be further enhanced while maintaining average response time targets. Our work can be improved in a number of ways. First, we further integrate load prediction method technique to fit our workload characteristics. Second, we will focus on expanding the utility analytic model to fit cloud environments with heterogeneous servers produced by different manufacturers. Third, we adopt Service Level Agreement (SLA) based negotiation of prioritized applications to determine the costs and penalties by the achieved performance level. If the entire request cannot be satisfied, some virtualized applications will be affected by their increased execution time, increased waiting time and or increased rejection rate.

### 7. References


Available: http://dx.doi.org/10.1007/3-540-44723-7| 2