A Survey on Adaptive resource provisioning for read intensive multi-tier applications in the cloud

1Manpreet Kaur, 2Dr. Rajiv Mahajan

1 M.Tech Student (CSE) GIMET, Amritsar
2 Professor in Dept. of CSE GIMET, Amritsar

ABSTRACT

With the growth of information technology, most important service is computing that meets everyday needs of general community. Cloud computing is that type of computing model which uses computing resources for delivering information technology services over a network. A Service-Level Agreement (SLA) guarantees for specific quality attributes to the consumers of services such as response time, throughput etc. This survey paper defines methodology of a prototype system which is used for automatic detection and resolution of bottlenecks in a multi-tier Web application hosted on a cloud in order to provide surety of maximum response time requirements. It also defines a method for identifying and retracting over provisioned resources in multi-tier cloud-hosted Web applications.

Keywords: Cloud computing, Adaptive resource management, Quality of service, Multi-tier applications, Service-Level Agreement.

1. INTRODUCTION

With the significant progress of Information Technology, Computing is one of the most important service that meets everyday needs of general community. Cloud computing is that type of computing model which uses computing resources for delivering information technology services over a network (typically the internet). Cloud providers [1] use the infrastructure as a service model to allow consumers to rent computational and storage resources on demand and according to their usage. Cloud infrastructure providers maximize their profits by fulfilling their obligations to consumers with minimal infrastructure and maximal resource utilization.

2. Adaptive Resource Management:

Adaptive Resource Management is also known as adaptive management. Adaptive management is a systematic iterative process for robust decision in face of uncertainty. Its main aim to reduce uncertainty over time via system monitoring. In this way, decision making simultaneously meets one or more resource management objectives and either passively or actively, acquires information needed to improve future management. Adaptive Management is based on learning process that’s why it improves long-run management outcomes.

3. Quality of services

The quality of service (QoS) refers various related aspects of computer networks and telephony that allow the transport of traffic with special requirements such as response time, jitter, latency, throughput etc.

4. Multi-tier applications

Multi-tier application architecture provides a model for developers which are used to create a reusable and flexible application. If developers want only to modify or add a specific layer, rather than have to rewrite the whole application over or if they decide to change technologies or scale up, they only breakup the application into tiers.

![N-Tier Application Architecture](image.png)

FIG. 1 N-TIER APPLICATION ARCHITECTURE
5. Service-Level Agreement
A service-level agreement (SLA) is basically a service contract where various services are formally defined. Sometimes the term SLA is used to refer to the contracted delivery time of the service or performance. As an example, internet service providers will commonly include service level agreements within the terms of their contracts with customers to define the level(s) of service being sold in plain language terms. In this case the SLA will typically have a technical definition in terms of mean time between failures (MTBF), mean time to repair or mean time to recovery (MTTR); various data rates; throughput; jitter; or similar measurable details.

6. RELATED WORK
The basic level of computing service that is considered[1] essential to meet the requirements of general community. For deliverer this vision various paradigms are proposed from which cloud computing is latest one, which also provide architecture creating Clouds with market-oriented resource allocation by leveraging technologies such as Virtual Machines (VMs). In [2]a methodology and presents a working prototype system is proposed for automatic detection and resolution of bottlenecks in a multi-tier Web application hosted on a cloud in order to satisfy specific maximum response time requirements. Also proposes a method for identifying and retracting over-provisioned resources in multi-tier cloud-hosted Web applications. In [3], paper consider multi-tier applications and propose the algorithm which is based on heuristics for identifying bottleneck problem. In this paper[4], EUCALYPTUS – an opensource software framework for cloud computing that implements what is commonly referred to as Infrastructure as a Service (IaaS). EUCALYPTUS enables users familiar with existing Grid and HPC systems to explore new cloud computing functionality while maintaining access to existing, familiar application development software and Grid middle-ware. This paper[5] is based on earlier work on dynamic CPU allocation to applications on shared servers, and present a feedback control system consisting of two nested integral control loops for managing the QoS metric of the application along with the utilization of the allocated CPU resource. The control system was implemented on a lab testbed using the Apache Web server as the application and the 90th percentile of the response times as the QoS metric. By testing the system using a synthetic workload based on an industry benchmark, we validate the two important features of the nested control design. First, compared to a single loop for controlling response time only, the nested design is less sensitive to the bimodal behavior of the system resulting in more robust performance. Second, compared to a single loop for controlling CPU utilization.
In this paper[6], an adaptive resource control system that dynamically adjusts the resource shares to individual tiers in order to meet application-level QoS goals while achieving high resource utilization in the data center. Control system is developed using classical control theory, and used a black-box system modeling approach to overcome the absence of first principle models for complex enterprise applications and systems. To evaluate controllers, a testbed simulating a virtual data center using Xen virtual machines was built. This paper evaluates[7] the overhead of a dynamic allocation scheme in both system capacity and application-level performance relative to static allocation. Experiments conducted with virtual containers built using Xen and OpenVZ technologies for hosting both computational and transactional workloads. In this paper, [8] descriptions and methods allow to deploy flexibly configured virtual cluster workspaces in the Grid. Performance evaluation results show that virtual clusters representing current Grid production environments can be efficiently deployed and managed and provide an acceptable platform for Grid applications. In an environment where both resource availability and software requirements evolve rapidly, this leads to resource underutilization, user frustration, and much wasted effort spent on bridging the gap between applications and resources.

7. Dynamic Provisioning for multi-tier Web applications
Dynamic provisioning takes place for multi-tier web applications by Reactive Model and Predictive Model which is used for Scale-up operation and Scale-down operation respectively. [2] A high-level flow diagram for bottleneck detection, scale-up decision making, and scale-down decision making in prototype system is [2] shown in Fig.2.

Fig.2 Flow Diagram for dynamic provisioning for multi-tier web applications that detect bottleneck and dynamically scales the tier to satisfy a SLA that defines response time requirements and ensures the release of over provisioned resources.
a) **Predictive Model for scale-up:** Reactive model is used for scale-up operation. When system starts it reads proxy logs for t seconds then it categorized requests such as dynamic request and static request and then calculates the response time of these requests. Now system checks whether last scale operation is realized. If yes then it again checks whether response time of static and dynamic request above the threshold value or not. If response time of static request is above from threshold value then it scales up the web server tier. Or If response time of dynamic request is above from threshold value then it gets CPU utilization of every instance in web server tier. Now systems checks is CPU in any instance in web tier reached saturated threshold If yes then it scale-up web tier or else scale-up database tier.

b) **Predictive Model for scale-down:** Predictive Model is used for scale down operation. For determine when scale down operation takes place here regression model is used that predicts each interval time, number of web server and database instances which is required for current workload. When response time of dynamic and static request is below than threshold value then through prediction model system gets number of web and database instances if web instances is more than current number of instances then it scale down web tier. If database instances is more than current number of database instances then scale down database tier.

8. CONCLUSION
With the significant progress of Information Technology, Computing is one of the most important service that meets everyday needs of general community. Cloud computing is that type of computing model which uses computing resources for delivering information technology services over a network. In this study, the prototype use, the heuristics for reactive scale-up of multi-tier Web applications, the predictive models for scale-down, and an evaluation of the prototype on a testbed cloud. There are a few limitations to this preliminary work. We only address scaling of the Web server tier and a read-only database tier. Our system only performs hardware and virtual resource management for applications. In particular, we do not address software configuration management; for example, we assume that the number of connections from each server in the Web server tier to the database tier is sufficient for the given workload.

REFERENCES