Resource Management on Clouds: the Multifaceted Problem and Solutions

Shikharesh Majumdar[1]

Real Time and Distributed Systems Research Centre,
Dept of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Canada
[1]majumdar@sce.carleton.ca

Extended Abstract

***Keywords*— resource management on clouds, cloud scheduling, matchmaking, resource management middleware, system performance**

1. Introduction

The popularity of cloud computing that provides resources on demand is spreading rapidly among service providers and consumers (clients). Techniques for performing efficient cloud computing have become a subject of interest to researchers as well as system builders. Virtualization is an important attribute of cloud computing. Virtualization can be performed at various levels: at the infrastructure level (infrastructure as a service) at the platform level (platform as a service) or at the software/application level (software as a service). Different vendors offer virtualized resources at one or more of these levels. The rise in interest in cloud computing can be inferred form the steady growth in the virtualization market. Both Meryl Lynch and Gartner have predicted a multi-billion dollar market for the cloud computing industry. As pointed out in [5], the proportion of annual world-wide enterprise IT spending on cloud computing in 2008 is expected to increase by more than double in 2012. A number of features of a cloud that has led to its popularity is described.

Dynamic change in resource requirements: A cloud lets the client's computing demand increase/decrease in accordance with its current requirements. Thus, temporary increases in resource usage for the service consumer can be effectively handled

Pay as you go: A cloud client does not need to own resources. The "pay as you go" model allows users to acquire resources on demand and use them only for the required period of time and pay a rental fee to the service provider. This is very useful for both small start up companies as well as larger enterprises. The smaller companies do not need to make a heavy investment on IT whereas the IT related costs are greatly reduced for the larger enterprises.

- *Energy savings:* Clouds are often considered to be in line with the initiatives *of green computing*. An efficient sharing of resources is achieved by consolidating the IT operations of multiple companies in a single data centre; this can potentially reduce the energy consumed by the computing, storage and cooling equipments.

Various types of clouds are described in the literature. These include *public clouds* offering resources that can be used by any consumer. The access of resources within private *cloud* is restricted to the members of a given group only. *Enterprise clouds* that serve a specific company and a *research and engineering* cloud that unifies resources including computing, storage devices as well software tools are examples of different types of the private cloud. In any of these cloud types, effective resource management is an important concern. A survey published in IDC identifies security and performance to be the two top concerns for cloud service consumers [4]. The resource management strategy deployed at a cloud is crucial for effectively utilizing the pool of resources and achieving a high system performance. As pointed out in the case of grids [3], quality of service (QoS) is an important issue in the context of clouds. *Service Level Agreement (SLA)*, an important characteristic of service requests submitted to a cloud [1] often requires the management of Advance Reservation (AR) requests that further complicates resource management and forms an important part of research on cloud resource management.

Although cloud computing provides a number of opportunities it also poses a number of challenges that include the following.

- *Security and privacy*: Both the execution of a client's application and the storage of client data are performed on a shared environment provided by the cloud. Appropriate mechanisms should be in place such that the results of application execution and the stored data must only be accessible to designated persons and machines identified by the client.
- *Heterogeneity and inter-operability*: Diversity in resources needs to be handled. This may include the diversity presented by devices ranging from various computing and storage devices to data analysis tools

This paper appeared in the Proceedings of the International Conference on Advances in Cloud Computing (ACC-2012), July 26-28, 2012 Bangalore, India organized by Computer Society of India, Editors: Rajiv Ranjan, Raj Buyya and Anirban Basu

that are unified by a research and engineering cloud or the variations in operating systems running on the servers in a data centre cloud

- *Lack of control on the execution environment*: The user's application is run on a system operated by the service provider and the user has little or no knowledge of the service provider's environment.
- *System monitoring*: Adequate facilities for monitoring system performance are required for ensuring that the desired levels of service are achieved.
- *Resource management*: Achieving the desired level of system performance is important for the client whereas revenue earned from the cloud is crucial for the service provider. System performance is often captured in terms of an SLA between the client and the service provider. Effective management of resources is needed for simultaneously achieving both the performance and revenue objectives of a system.

This paper focuses on *resource management* on clouds. Some of the other issues discussed earlier are also referred to in the context of resource management that forms the main direction of the discussion presented.

2. Resource Management

The important operations performed by a resource manager in a cloud includes: *matchmaking* and *scheduling*. Given a pool of resources a matchmaking algorithm chooses the resource or resources to be allocated to an incoming client request. Once a number of requests get allocated to a specific resource, a scheduling algorithm is used to determine the order in which these requests are to be executed for achieving the desired system objectives. Both matchmaking and scheduling are hard problems because they need to satisfy user requirements for a quality of service defined in a service level agreement while generating a high resource utilization and/or adequate revenue for the service provider. In some systems both the scheduling and matchmaking operations are combined and performed in one resource management step. A resource management framework for grids that includes both matchmaking and scheduling described in [2] is being extended to clouds.

User requests handled by a resource manager are often characterized as on demand and advance reservation.

Advance Reservation Request (AR): is characterised by an earliest start time, a request execution time and a deadline that are specified by the user. By accepting such a request, a service provider enters into a service level agreement with the client.

On Demand Request: An on demand request is not associated with a deadline and is satisfied on a best effort basis.

Note that other types of SLAs regarding system security and availability are not considered in this paper that primarily focuses on the handling of ARs and ODs by the resource manager.

Both matchmaking and scheduling are discussed in the context of systems subjected to ODs and ARs. The job of making an effective matchmaking and scheduling decision is further complicated by a number of factors that include the following:

- Coscheduling of resources: Certain applications may require multiple resources to be allocated at the same time. Effective techniques for resource coscheduling need to be deployed for handing such applications.
- Handling Uncertainty: Uncertainties associated with request execution times and the local resource management policy used at a specific resource increases the complexity of resource management.

A. Uncertainty associated with request execution times

Estimating the execution time for an application request that needs to be specified in an AR is a non-trivial task. Existing studies in the literature show that application execution times are often grossly overestimated. Two techniques, one that artificially adjusts the request execution times [2] and one that performs resource "over booking" [2] will be discussed. Although [2] focuses on grids the techniques described can be easily extended to clouds. Another technique that introduces the notion of *soft SLAs* that allow a certain proportion of ARs to miss deadlines is currently under investigation will also be described. Issues such as system monitoring and security that are related to resource management will be discussed

B. Lack of a priori knowledge of Local Resource Management Policy

In a cloud comprising a number of heterogeneous resources each resource is often subjected to its own local scheduling policy. The scheduling policy deployed at each resource may be unknown to the resource broker performing matchmaking. This is because the system configuration for a cloud may not be completely known at the time of system design or the resources included in the cloud may change during the lifetime of the system**.** An a*ny schedulability criterion* is proposed in [6] for handling such an environment. This criterion comprises a set of inequalities characterized by the attributes of the unfinished requests on the system that are evaluated when a new request arrives. If all the inequalities are satisfied at a resource, the incoming request is guaranteed to meet its deadline irrespective of the scheduling policy used at the resource; only the assumption that a *work-conserving* scheduling algorithm used at the resource needs to be satisfied. Matchmaking techniques that use this any-schedulability criterion are described in [7].

This paper appeared in the Proceedings of the International Conference on Advances in Cloud Computing (ACC-2012), July 26-28, 2012 Bangalore, India organized by Computer Society of India, Editors: Rajiv Ranjan, Raj Buyya and Anirban Basu

The resource management techniques discussed are often based on multiple objectives. These include the satisfying of SLAs as captured in an advance reservation request, providing satisfactory response times to on demand requests as well as generating ample revenue for the service provider.

C. Real Time Issues

Real time issues need to be handled in the context of certain clouds. A research and engineering cloud for sensor-based bridge infrastructure management is currently being investigated at Carleton University. Management of bridges in most developed nations is an expensive undertaking. One of the problems with the current approach to bridge infrastructure management is that each bridge owner maintains its infrastructure in isolation from other stakeholders in a city, province or country. Unification of resources available at each of these stakeholders and making them available on demand can greatly reduce this maintenance cost as well as reduce the number of accidents that occur from failing infrastructures. A cloud that can unify various resources including computing and storage resources, archival databases, sensor data repositories, software tools for data analysis as well human experts and make them available on request to the cloud user is currently being investigated. Emergency response required for handling emergencies occurring on a bridge is an important concern. Effective scheduling and matchmaking techniques that can handle resource requests submitted during an emergency are an important component of the collaborative project supported by Cistel Technology and the Ontario Centres of Excellence (OCE) in Canada.

3. Summary

Resource management is of critical importance in the context of clouds. The various facets of resource management make it an interesting topic worthy of research. A number of important issues related to resource management were described earlier. These include resource scheduling and matchmaking algorithms in general and techniques for coscheduling of resources and for handling various uncertainties associated with the workload and the local resource scheduling policies in particular. Additional issues such as the interaction of system security with resource management and performance warrant investigation. Resource management in clouds often have multiple objectives: achieving high system performance, generating ample revenue for the service provider as well as meeting service level agreements between the client and the service provider. The additional objective of achieving savings in energy is providing new challenges for system researchers and builders.

Acknowledgments

Research results from a number of research projects are described. Supports from Nortel, Cistel Technology, Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Centres of Excellence are gratefully acknowledged.

References

[1] Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I., "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility", in *Future Generation Computer Systems* 25 (2009), 2009, pp. 599-616.

[2] Farooq, U., Majumdar, S., Parsons, E., "Achieving Efficiency, Quality of Service and Robustness in Multi-Organizational Grids", in *Journal of Systems and Software* (*Special Issue on Software Performance*), 82(1), January 2009, pp. 23-38.

[3] I. Foster, C. Kesselman, S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," in *Int'l Journal of Supercomputer Applications*, 15(3), 2001.

[4] Gens, Frank, "IT Model in the Cloud Computing Era", in *IDC Enterprise Panel*, August 2008.

[5] Gens, Frank, "IT Cloud Services Forecast – 2008, 2012: A Key Driver of New Growth" in *IDC Exchange*, October 2008.

[6] Majumdar, S. "The "Any-Schedulability" Criterion for Providing QoS Guarantees Through Advance Reservation Requests", in the *Proceedings of the Cluster Computing and the Grid* (*International Workshop on Cloud Computing)*, Shanghai (China), May 2009, pp. 490-495.

[7] Melendez, J.O., Majumdar, S., "Matchmaking on Clouds and Grids", *International Journal of Internet Technology (JIT)* (accepted for Publication), 2012.