International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Dynamic Resource Allocation for Cloud Computing Including the Feasibility Study on Green Computing (Using Need Based Resource Allocation Scheme)

B. Karthikeyan[1], T. Sasikala[2] , M. DilliBabu[3]

[1]Research Scholar Anna University, Chennai, India

[2]Principal, SRR Engineering College, Chennai, India

[3]Assistant Professor, Panimalar Engineering College, Chennai, India

Abstract: ***Computing technology allows the user to avail all the feasible usage extract from it, in which cloud computing explicitly shows the utilization of resources based on the needs of business customers as well as the sophisticated users. Many of the utilized gains in the model demonstrate the user about the resource usage from different sources and coordinate them in to gained resources through the amalgamation of resources in the server. In this paper we propose a new technology named "Need Based Resource Allocation (NBRA)" which allocates resources whenever needed with maximum utilization. Skewness is used to find out the eventuality in the resource allocation in the server side and resource utilization in the client side, for both the criteria, many dimensional views with the prediction are carried out to prove the same concept. To minimize the wastage of resources, we first check the proposed workload to that of the existing scenario. Before allocation of resources' to an event, we have the confirmed resource required to complete the event and we make sure that no resources are over allocated and under-allocated. Thus our paper confirms that dynamic allocation of resources will not compromise the performance of the process.***

Keywords: Skewness, Need Based Resource Allocation, dynamic resource allocation, over allocated and Under Allocated

1. Introduction

1.1 Motivation

The growing popularity of the World Wide Web has led to the advent of Internet data centers that host third-party web applications and services. A typical web application consists of a front-end web server that services HTTP requests, an application server that contains the application logic, and a backend database server. In many cases, such applications are housed on managed data centers where the application owner pays for server resources, and in return, the application is provided guarantee on resource availability and performance. To provide such guarantees, the data center— typically a cluster of servers—must provide sufficient resources to meet application needs. Such provisioning can be based either on a dedicated or a shared model.

In the dedicated model, some number of cluster nodes are dedicated to each application and the provisioning technique must determine how many nodes can be allocated to the application. In the shared model, an application can share node resources with other applications and the provisioning technique needs to determine how to partition resources on each node among the successive applications**,** which results the system more complex. The fraction of the resources allocated to each application depends on the expected workload which fulfils the system requirements. The workload of web applications is known to vary dynamically over multiple time scales and it is challenging to estimate such workloads with respect to Apriori. Consequently, static allocation of resources to applications is problematic—providing more resources based on worst case workload estimates, can result in potential wastage of resources, where as allocating fewer resources can result in violation of guarantees. An alternate approach is to allocate resources to applications based on the variations in their workloads. In this approach, each application is given a certain minimum share based on coarse-grain estimates of its resource needs; the remaining server capacity is dynamically shared among various applications based on their instantaneous needs.

To illustrate, consider two applications that share a server and are allocated 30% of the server resources each; the remaining 40% is then dynamically shared at run-time so as to meet the guarantees provided to each application. Such resource sharing can yield potential multiplexing gains, while allowing the system to react to unanticipated increases in application load and thereby meet QoS guarantees. Resource allocation method that can handle changing application workloads in shared data centers is the focus of this paper.

2. Existing Work

A growing number of companies have to process huge amounts of data in cost-efficient manner. Classic representatives for these companies are operators of Internet search engines. The vast amount of data they have to deal with every day has made traditional database solutions prohibitively expensive.

Instead, these companies have popularized an architectural paradigm based on a number of commodity servers. Problems like processing crawled documents or regenerating a web index are split into several independent subtasks, distributed among the available nodes, and computed in parallel.

2.1 LIMITS

The cloud's virtualized nature helps to enable promising new use cases for efficient parallel data processing.
1. However, it also imposes new challenges compared to classic cluster setups.
2. The major challenge we see is the cloud's opaqueness with prospect to exploiting data locality.

3. Literature Survey

In the related study, they focused on bag of tasks workload type and propose an idea to facilitate dynamic resource allocation for it. Technically, the proposed approach exploits users' service level agreement parameters and classifies them. It controls utilization of servers to response users in a reasonable time the proposed approach uses a network based approach by exploiting two main parts, namely classifier and Resource Allocator (RA). Users' requests as the preliminary actor receive by the classifier. It checks clients' requests. After that, the RA provides initializing process for the BoTs to run them based on available resources.

In the Efficient and Trustworthy Resource Sharing Platform for Collaborative Cloud Computing proposed about the opportunities and challenges for efficient parallel data processing in clouds and present our research project Nephele. Nephele is the first data processing framework to explicitly exploit the dynamic resource allocation offered by today's IaaS clouds for both, task scheduling and execution. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution. Based on this new framework, we perform extended evaluations of MapReduce-inspired processing jobs on an IaaS cloud system and compare the results to the popular data processing framework Hadoop Nephele's architecture follows a classic master-worker pattern. Before submitting a Nephele compute job, a user must start a VM in the cloud which runs the so called Job Manager (JM). The Job Manager receives the client's jobs, is responsible for scheduling them, and coordinates their execution. It is capable of communicating with the interface the cloud operator provides to control the instantiation of VMs. We call this interface the Cloud Controller. By means of the Cloud Controller the Job Manager can allocate or deallocate VMs according to the current job execution phase.

Data Center Resources for Cloud Computing presents an vision, Architectural Elements, and Open Challenges for energy-efficient management of Cloud computing environments. We focus on the development of dynamic resource provisioning and allocation algorithms that consider the synergy between various data center infrastructures (i.e., the hardware, power units, cooling and software), and holistically work to boost data center energy efficiency and performance. In particular, this paper proposes (a) architectural principles for energy-efficient management of Clouds; (b) energy-efficient resource allocation policies and scheduling algorithms considering quality-of-service expectations, and devices power usage characteristics; and (c) a novel software technology for energy-efficient management of Clouds.

The main objective for the work is to initiate research and development of energy-aware resource allocation mechanisms and policies for data centers so that Cloud computing can be a more sustainable eco-friendly mainstream technology to drive commercial, scientific, and technological advancement for future generations. Specifically, our work aims to:

- Define an architectural framework and principles for energy-efficient Cloud computing;
- Investigate energy-aware resource provisioning and allocation algorithms that provision data center resources to client applications in a way that improves the energy efficiency of the data center, without violating the negotiated Service Level Agreements (SLA);
- Develop autonomic and energy-aware mechanisms that self-manage changes in the state of resources effectively and efficiently to satisfy service obligations and achieve energy efficiency;
- Investigate heterogeneous workloads of various types of Cloud applications and develop algorithms for energy-efficient mixing and mapping of VMs to suitable Cloud resources in addition to dynamic consolidation of VM resource partitions; and
- Implement a prototype system – incorporating the above mechanisms, and techniques – and deploy it within the state-of-the-art operational Cloud infrastructures with real world demonstrator applications.

4. Proposed Work

In recent years a variety of systems to facilitate MTC has been developed. Although these systems typically share common goals (e.g. to hide issues of parallelism or fault tolerance), they aim at different fields of application.

Map Reduce is designed to run data analysis jobs on a large amount of data, which is expected to be stored across a large set of share-nothing commodity servers. Once a user has fit his program into the required map and reduce pattern, the execution framework takes care of splitting the job into subtasks, distributing and executing them. A single Map Reduce job always consists of a distinct map and reduce program. we propose a new technology named "Need Based Resource Allocation (NBRA) " which allocates resources whenever needed with maximum utilization. Skewness is used to find out the eventuality in the resource allocation in the server side and resource utilization in the client side, for both the criteria, many dimensional views with the prediction are carried out to prove the same concept. To minimize the wastage of resources, we first check the proposed workload to that of the existing scenario. Before allocation of resources' to an event, we have the confirmed

resource required to complete the event and we make sure that no resources are over allocated and under-allocated. Thus our paper confirms that dynamic allocation of resources will not compromise the performance of the process.

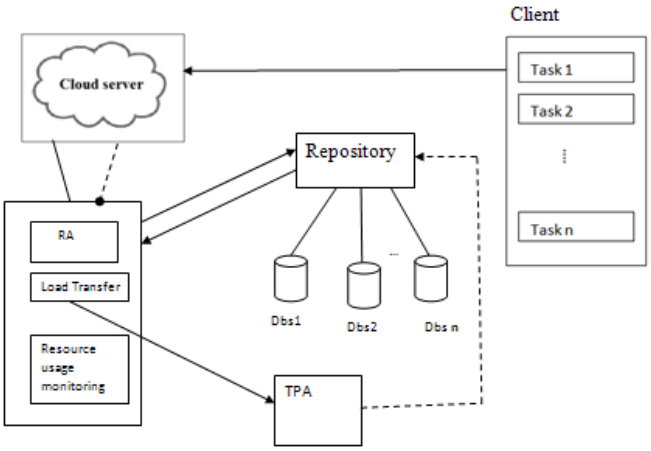

Figure 1: Proposed Architecture for Need Based Resource Allocation Method

The proposed system enumerates the features of allocating the resources from the repository in a dynamic manner for the scheduled task. The repository contains the various databases in an orderly manner. Now, if the client has been assigned with different tasks, the tasks moves to the cloud server which are accompanied by the repository. The resource allocator in the server act as an interface between the task to that of the repository, once the communication is established between the repository and RA, the Load Transfer initiates its dynamic allocation of resources to the tasks under the supervision of TPA. TPA monitors only the load transfer for the desired tasks and not acknowledges the process status. The process status with respect to the resources is achieved by the Resource Usage monitoring and thus confines the Dynamic allocation of resources to the Scheduled tasks without any intervention.

4.1 Advantage

- Takes less Time for the Data Transfer.
- Effective and Speedy data Transfer over the network.
- Takes less bandwidth as we compress the files.

5. Conclusion & Future Work

In this paper we have discussed the challenges and opportunities for efficient parallel data processing in cloud environments and presented the new algorithm for dynamic resource allocation in cloud computing environment using "Need Based Resource Allocation". This algorithm allocates the resources for the tasks in dynamic manner with maximum utilization of resources and thus meets the required QoS to the users. We conclude that no resources are over utilized and underutilized using our proposed method. As future work, we plan to add more security with our existing work. And also to integrate the mobile environment where it requires optimum resource utilization.

References

[1] http://aws.amazon.com, 2013.
[2] Saurabh Kumar Garg and Rajkumar Buyya, "Green Cloud computing and Environmental Sustainability".
[3] Dropbox, www.dropbox.com, 2013.
[4] Daniel Warneke and Odej Kao, "Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud", ieee transactions on parallel and distributed systems, vol. 22, no. 6, June 2011.
[5] Chun-Cheng Lin, Hui-Hsin Chin, and Der-Jiunn Deng "Dynamic Multiservice Load Balancing in Cloud-Based
[6] Multimedia System" ieee systems journal, Vol.8, no. 1, March 2014
[7] Rajkumar Buyya, Anton Beloglazov, and Jemal Abawajy " Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges" Proceedings of the 2010 International Conference on
[8] Parallel and Distributed Processing Techniques and Applications (PDPTA 2010), Las Vegas, USA, July 12-15, 2010
[9] Zhongjie Wang · Xiaofei Xu "A sharing-oriented service selection and scheduling approach for the optimization of resource utilization" Springer-Verlag London Limited 2011.
[10] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. Sixth Conf. Symp. Opearting Systems Design and Implementation (OSDI '04), p. 10, 2004.
[11] A.Suphalakshmi and Sreejith M, "An intelligent, energy conserving load balancing algorithm for the cloud environment using ant's stigmergic behavior", International Journal of Communications and Engineering Volume 04- Issue: 03March2012.
[12] Er. Navdeep Kochhar and Er. Arun Garg,"Eco-friendly computing: green computing",international journal of computing and business research issn (online) : 2229-6166.volume 2 issue 2 may 2011.