Efficient Parallel Data Processing in the Cloud

THANAPAL.P
Assistant Professor (senior)
School of Information Technology and Engineering
VIT University
Vellore, India
thanapal_mdu@yahoo.com

NISHANTHI.S.P
School of Information Technology and Engineering
VIT University
Vellore, India
nishanthipichandi@yahoo.com

Abstract:

Cloud computing is a distributed computing technology which is the combination of hardware and software and delivered as a service to store, manage and process data. A new system is proposed to allocate resources dynamically for task scheduling and execution. Virtual machines are introduced in the proposed architecture for efficient parallel data processing in the cloud. Various virtual machines are introduced to automatically instantiate and terminate in execution of job. An extended evaluation of MapReduce is also used in this approach.

Keyword - cloud computing, parallel, data processing.

I. Introduction:

In recent years many number of growing companies have a problem in processing large amount of data in cloud in low cost. They used many number of commodity servers for an architectural paradigm to process large amount of data for many growing companies[1]. The large amount of data is divided into several independent sub tasks, distributed among nodes that are available and compute them in parallel.

Customized processing frameworks are built by many companies instead of developing distributed applications and simplify their complexity.They are classified into two terms, high throughput computing and many-task computing. These are classified based on the amount of data and number of jobs involved in the particular task execution. There are different types of virtual machines depends on the number of CPU cores and amount in main memory[8]. The virtual machines charge the companies only the amount of time they are used.

In this paper, we going to split the tasks into subtasks and they are assigned to different employees in the team and the resources are assigned dynamically using virtual machine. This virtual machine will start and terminate as the resources needed for the tasks that going to be executed[3]. The strategies for scheduling are used in the framework. The scheduling concepts give easy method to use the resources and process the data in parallel. This framework will dynamically allocate the resources as the tasks needed them[1]. Current data processing framework are required the cloud to be static in cluster environments, this leads to big drawback in allocating dynamic resources. To overcome certain problem, the virtual machine is used in the framework of processing data in parallel.

II. Related Work:

Daniel Warneke and Odej[1] Kao proposed a parallel data processing in the cloud using nephele architecture where a virtual machine is installed for execution of job description and they are charged and assigned automatically when the resources are needed.Chao-Rui Chang , Meng-Ju Hsieh, Jan-Jan Wu, Po-Yen Wu, Pangfeng Liu [2] proposed a new distributed B-tree column indexing scheme for HBase, which support indexing for non-row-key columns as well as parallel B-tree search in large data table. Their results show both performance and scalability advantage of indexing schemes on point queries, range queries and aggregation operations compared with HBase. Qingjia Huang, Sen Su, Jian Li, Peng Xu, Kai Shuang, and Xiao Huang[3] proposed an approach called enhanced energy-efficient scheduling(EES) algorithm to reduce the energy consumption while meeting the performance -based service level agreement(SLA). They exploit the slack room and allocate the non-critical Jobs in a global manner in their schedule. Their results shows that EES is able to reduce considerable energy consumption because of using random generated and real-life application workflows. Muhammad Baqer Mollah, Kazi Reazul Islam*, Sikder Sunbeam Islam [4] presents all the cloud computing architectures, advantages, platforms, issues and challenges, applications, future and research options. There are four types of generations in cloud computing, they are mainframe based computing , personal computing, client server based computing, web server based computing. They results in several advantages such as fast micro-processor, huge memory, high-speed network, reliable system architecture, etc. Ting He, Shiyao Chen†, Hyoil Kim‡, Lang Tong†, and Kang-Won Lee[5] proposed an efficient heuristic scheduling policy by formulating the problem as restless Multi-Armed Bandits(MAB) under relaxed synchronization. They prove the indexability of the problem and provide closed-form formulas to compute the indices. The proposed index policy achieves consistently good performance under various server dynamics compared with the existing policies. E. Atanassov, T. Gurov and A. Karaivanova [6] proposed steps and best practices that could decrease the risk from penetration by random as well as determined attackers and outline services that could bring down the overall security risks.H. Topcuoglu, S. Hariri, and M. Wu[7] describes a well-known heuristic algorithm heterogeneous earliest finish time (HEFT). Various heuristics are proposed and that are differentiated as list-scheduling algorithms, cluster-based algorithms, duplication- based algorithms and guided random search methods. Heuristics achieve good performance with low time complexity.M. Mezmaz, N. Melab, Y. Kessaci, Y. Lee, E. Talbi, A. Zomaya, andD. Tuyttens [8] presented a energy efficient scheduling. A genetic algorithm to balance the tradeoff performance and the energy consumption. They analyze and utilize the slack in a global manner and devise a scheduling heuristic algorithm to minimize the energy consumption while meeting deadline in heterogeneous computing system.M. Pokharel, YoungHyun Yoon, Jong Sou Park [9] cloud computing is differentiated into three segments such as applications, platforms and infrastructures. Each segment offers different products for businesses and also individuals around the world and serves for different purpose.

III. Objective:

The main objective is to dynamically allocate resources using virtual machines for executing large amount of data for efficient data processing in the cloud.

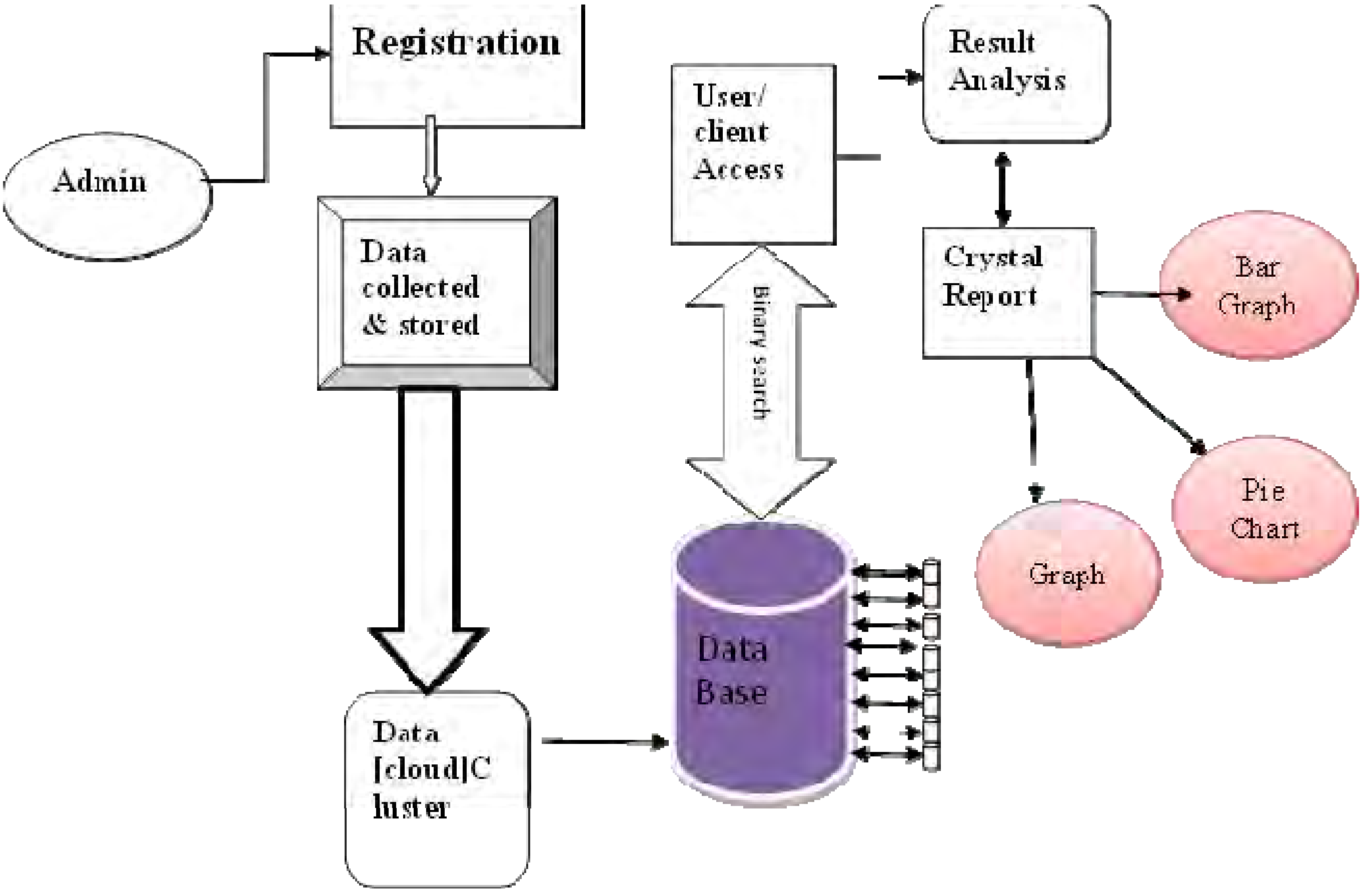


Fig 1. A proposed architecture for parallel data processing

IV. Problem Definition:

Server-client computing is a distributed architecture which partitions the tasks among service providers and service requesters[4]. A server is a host which performs many server programs and shares their resources whereas clients initiate the communication with the server and also share resources.

Administration part in the architecture diagram stores all the information regarding the task manager, job manager, team member and client who make their request to perform the scheduling and execution process[3]. Once the client registered in the administration, then all the details about their job and how and to whom to they are scheduled will be stored in administration, cloud and database.

By using location based services, the client may create a duplicate ID of their own. When a user wishes to make a query, her location will be sent to the trusted server through a secure connection[1]. The particular location will be hidden by Anonymizer Spatial Region(ASR) then the location's candidate set will be retrieved and query is passed.

To guarantee the k-anonymity in road networks, edge ordering anonymization approach is proposed for the users. Spatial cloaking techniques are used to prevent identity inference in location based services, which offers privacy protection and the current user cannot be identified or distinguished among them[5]. The most widely used model, k-anonymity is used for location based services.

The main idea of k-anonymity is to fake the user's exact location by using anonymizing spatial region (ASR). The K-anonymity is widely used method for privacy in location but, there are many location based attacks are possible.

A. Algorithms:

1.) Job Scheduling and Execution:

Once the valid job graph is received from the user, the job manager present inside the administration will convert the job graph into execution graph. An execution graph is the basic structure in scheduling and monitoring the job that is being executed[1]. It contains every information to schedule and execute job on the cloud.

2.) Parallelization and scheduling Strategies:

The channels present in the network are called network channels. Each vertex in the job graph is converted into execution vertex of execution graph. The user gives any job annotation and the execution vertex is assigned to its own instance. This method is continued until they are prohibited by scheduling restrictions.

B. Experiments:

1.) The first experiment reads the input data and sorts them in ascending order by giving integer numbers.

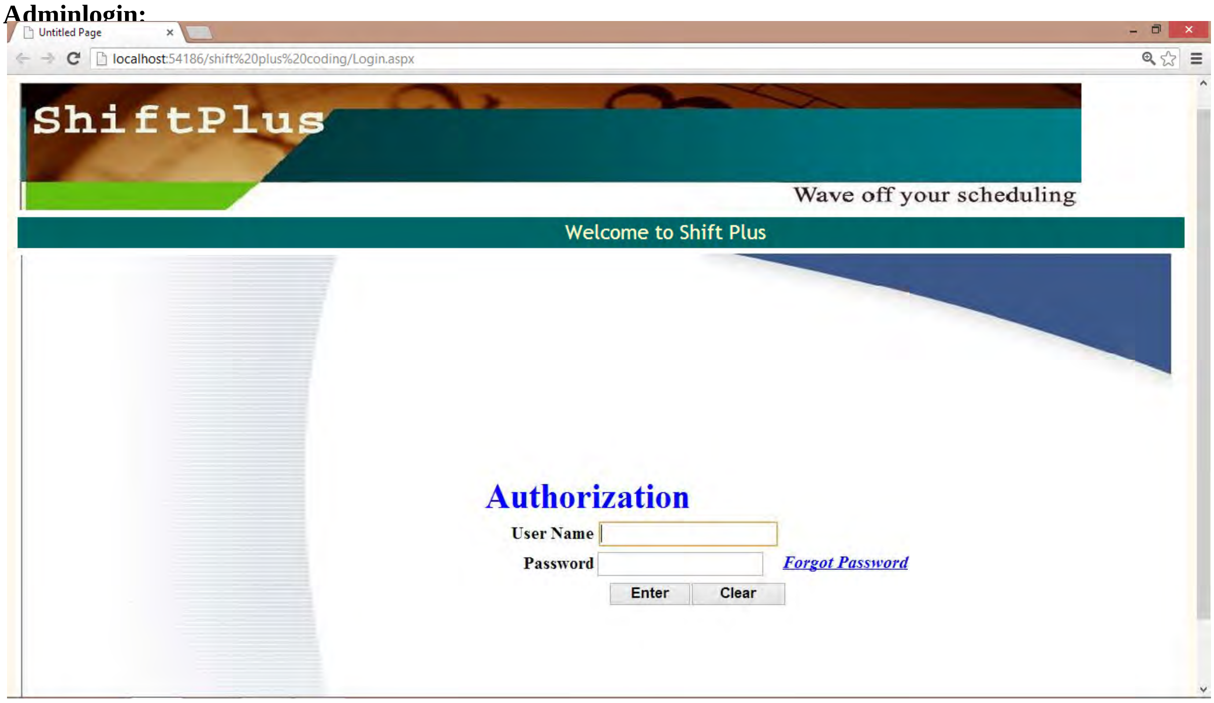
2.) We reused the same programs written for experiment one by using wrapper classes, which allows us to run the programs without modification.

3.) Sorts and aggregation maps are used instead of execution programs, to manage heterogeneous computed resources.

V. Result and Analysis:

We assigned the number of maps to be 48 and number of reducers to be 12, then there is improvement in memory file system to 1 GB and 512 MB. By using executed programs, the dynamic allocation and tasks are assigned by data flow control. The utilization has been increased from 60 to 80 % of resources and memory.

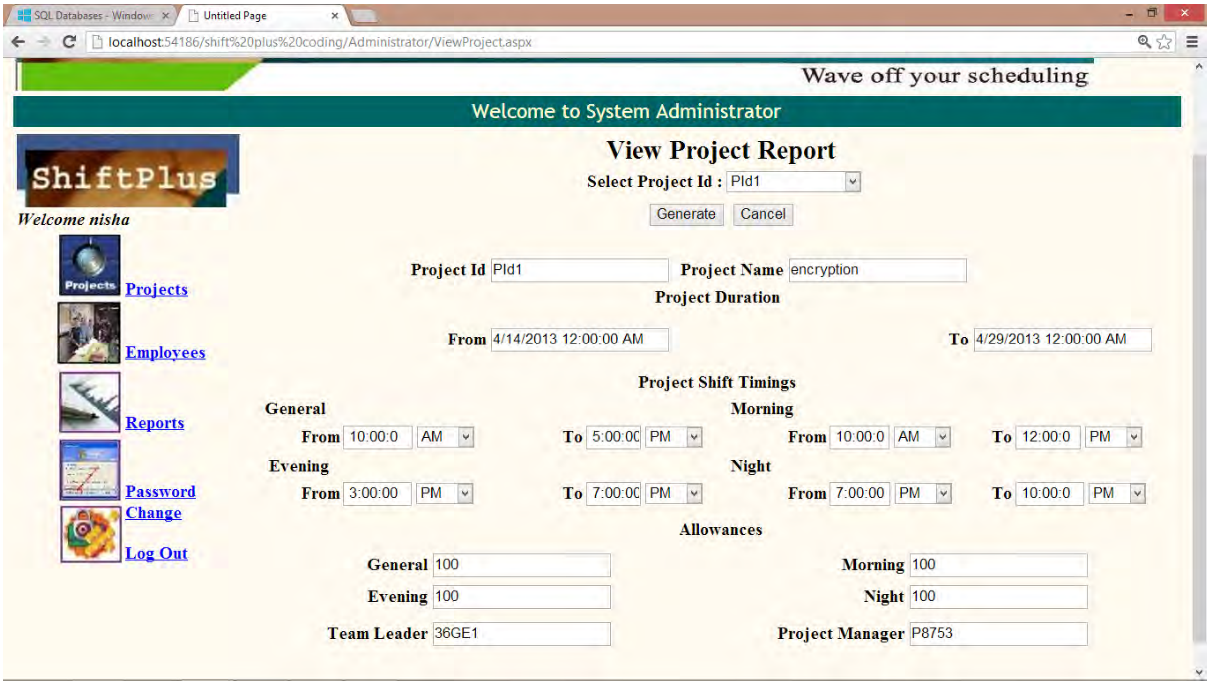
Adminlogin:



User Name

Password

project report details:



View Project Report

Select Project Id : PId1

Project Id PId1 Project Name encryption

Project Duration

From 4/14/2013 12:00:00 AM To 4/29/2013 12:00:00 AM

Project Shift Timings

General: From 10:00:0 AM To 5:00:00 PM

Morning: From 10:00:0 AM To 12:00:0 PM

Evening: From 3:00:00 PM To 7:00:00 PM

Night: From 7:00:00 PM To 10:00:0 PM

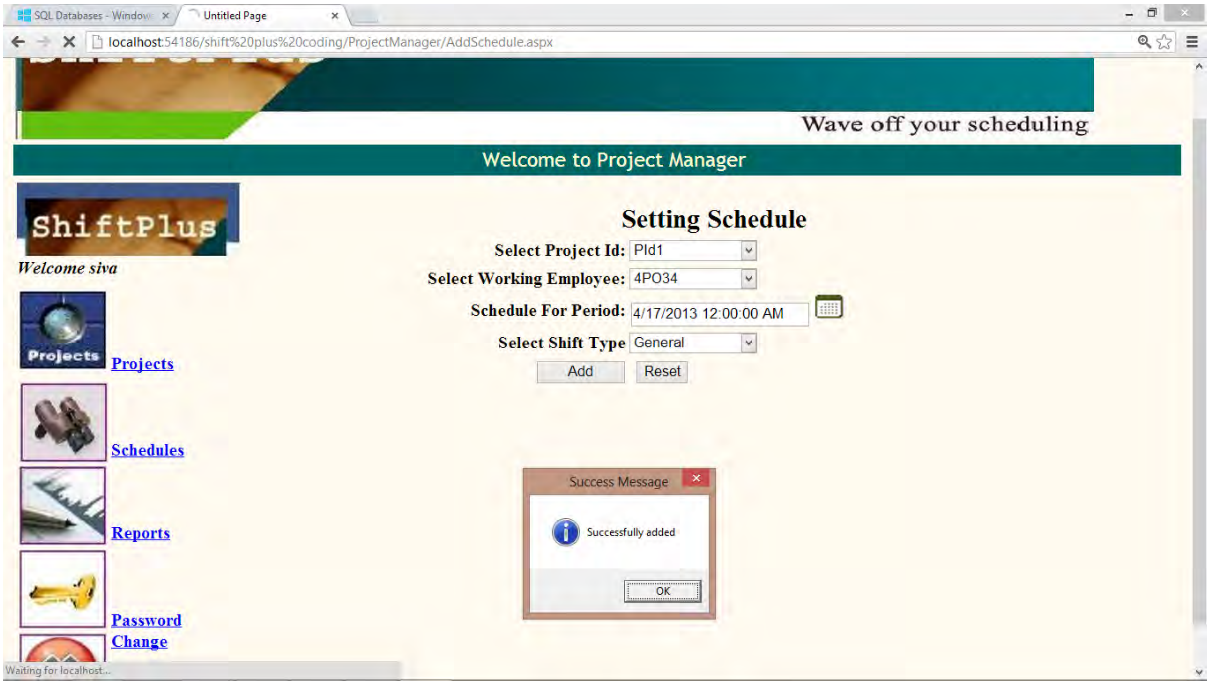
Allowances

General 100 Morning 100

Evening 100 Night 100

Team Leader 36GE1 Project Manager P8753

Schedule details:



Setting Schedule

Select Project Id: PId1

Select Working Employee: 4PO34

Schedule For Period: 4/17/2013 12:00:00 AM

Select Shift Type General

Successfully added

VI. Conclusion:

In this paper, we have seen how the resources are dynamically allocated using virtual machines and reduce the cost. The virtual machine is used to allocate the resources dynamically and schedule them in order to reduce the cost. The job execution in virtual machine can help the overall resource allocation and utilization. The virtual machine's ability can be improved the automatic allocation and underutilization.

VII. References:

[1] Daniel Warneke and Odej Kao, Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud, IEEE transactions on parallel and distributed systems, january 2011

[2] Chao-Rui Chang, Meng-Ju Hsieh, Jan-Jan Wu,HSQL: A Highly Scalable Cloud Database for Multi-User Query Processing, IEEE Fifth International Conference on Cloud Computing, pages 943-944, 2012.

[3] Qingjia Huang, Sen Su, Jian Li, Peng Xu, Kai Shuang, and Xiao Huang, Enhanced Energy-efficient Scheduling for Parallel Applications in Cloud, 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pages 781-786, 2012.

[4] Muhammad Baqer Mollah, Kazi Reazul Islam*, Sikder Sunbeam Islam,Next Generation of Computing through Cloud Computing Technology, 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE),2012.

[5] Ting He∗, Shiyao Chen†, Hyoil Kim‡, Lang Tong†, and Kang-Won Lee∗, Scheduling Parallel Tasks onto Opportunistically Available Cloud Resources, IEEE Fifth International Conference on Cloud Computing, pages 180-187, 2012.

[6] E. Atanassov, T. Gurov and A. Karaivanova, Security issues of the combined usage of Grid and Cloud resources,MIPRO 2012, pages 417-420,May 21-25,2012.

[7] H. Topcuoglu, S. Hariri, and M. Wu, "Performance-effective and lowcomplexitytask scheduling for heterogeneous computing," Parallel andDistributed Systems, IEEE Transactions on, vol. 13, no. 3, pp. 260–274,2002.

[8] M. Mezmaz, N. Melab, Y. Kessaci, Y. Lee, E. Talbi, A. Zomaya, andD. Tuyttens, "A parallel bi-objective hybrid metaheuristic for energyawarescheduling for cloud computing systems," Journal of Parallel andDistributed Computing, 2011.

[9] M. Pokharel, YoungHyun Yoon, Jong Sou Park, "Cloud Computing in SystemArchitecture", Proceedings of IEEE International Symposium on ComputerNetwork and Multimedia Technology, pp. 1-5, 2009.