Chapter 18
Dynamic Task Assignment with Load Balancing in Cloud Platform

Subhadarshini Mohanty
College of Engineering & Technology, India

Prashant Kumar Patra
College of Engineering & Technology, India

Subasish Mohapatra
National Institute of Technology, India

ABSTRACT
Load balancing is one of the major issue in cloud computing. Load balancing helps in achieving maximum resource utilization and user satisfaction. This mechanism transparently transfer load from heavily loaded process to under loaded process. In this paper we have proposed a hybrid technique for solving task assignment problem in cloud platform. PSO based heuristic has been developed to schedule random task in heterogeneous data centres. Here we have also used variants of Particle Swarm Optimization (PSO) which gives better result than PSO and other heuristics for load balancing in cloud computing environment.

INTRODUCTION
Rapid development in science & technology leads to the upgradation in processor, network, storage and computing resources. They have become more powerful and ubiquitous. This technological advancement ensures a new computing paradigm called cloud computing (Zhang et.al. 2010). It provides abundant of opportunities for both cloud service provider as well as cloud subscriber in the form of Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (Saas). On demand service, rapid elasticity, scalability, and metered service are major characteristics of clouds computing. Cloud provide services in different models such as SaaS, PaaS, IaaS (Sajid et.al. 2013). In SaaS (Software as a Service) cloud service providers update their application on the portal which can be accessed by various
subscribers. They do not worry about the infrastructure. Different applications are arranged in a unified logical environment on SaaS to provide optimal speed, availability, security and recovery etc. Consumers develop new cloud application and services directly on PaaS (Platform as service). It offers a fully developed platform that host both complete and in progress cloud platform (Sajid et.al. 2013). Cloud consumers directly use storage, network, processor provided in IaaS (Infrastructure as a Service) cloud.

Cloud provide large volume of data and files available for serving a number of users around the world. Managing such high volume of data require several methodology to optimize operations and provide better quality of service. It is essential to improve storage utilization and the download performance for users. One important issue associated with this field is dynamic load balancing or task scheduling. Cloud datacenters spread in different geographical channels. Users can subscribe cloud service from any location. Random generation of task create load imbalance in cloud platform i.e some of the datacenters are heavily loaded while others are in idle or lightly loaded. Load measurement, Information exchange, load balancing operation are the major component of developing an efficient load balancing algorithm. As the load change in cloud platform dynamically, it is difficult to calculate load on individual datacenters efficiently (Dillon et al., 2010).

Various challenges come out at the time of designing load balancing algorithm such as distribution of datacenters in cloud platform, storage structure, replication and single point of failure. Some algorithms are designed to be efficient only for closely located nodes, where the communication delay is negligible. Similarly load balancing algorithm based on reliability which can overcome single point of failure but it include extra responsibility for managing the resources. Load balancing algorithms are basically classified into two different categories i.e. static algorithm and dynamic algorithm. Static algorithms have the prior knowledge about the node statistics and user requirements but dynamic load balancing algorithms are characterised by run time statistics of each node and are monitored to adopt changing load (Sinha, 1998).

Static algorithms are classified into two different categories such as deterministic and probabilistic. In deterministic the algorithms use the information about the properties of the nodes and the characteristics of the processes to be scheduled to deterministically allocate process to nodes. Probabilistic algorithm uses the information regarding the static attributes such as number of nodes, processing capability and the network topology.

Further dynamic algorithms are classified into two different categories such as centralized and distributed. In centralised algorithms single node is responsible for maintaining the statistics of entire work and update characteristics time to time. Decisions are made at single node which is called as centralised server node. It finalizes about the assignment of new process using the state information stored in it. This algorithm can efficiently make process assignment because the centralised server knows both the load and number of process needing services. The major drawback of this approach is associated with single point of failure (SPOC). So distributed algorithms are best alternate solution. Distributed algorithms operate on behaviour of all processors. All the processors are responsible for load balancing and storing the state information in local server. The classification of load balancing algorithm are depicted in Figure 1.

In on demand service the load information only be collected just making a load balancing decision. This method can minimize the overall communication cost. The processor will broadcast its load information to the distributed system when its load status changes. Large overhead in communication makes this rules impractical. So centralized algorithms can be used to achieve the above goal by allocating dedicated processor to collect and maintain the systems load information.