ABSTRACT

Request scheduling has a major impact on the performance of the service processing design in a large scale distributed computing environment. Adopting a suitable request scheduling policy helps to achieve high performance. It is desirable in a distributed system to have a request scheduling principle that evenly distributes the workload among the servers, according to their capacities. The capacities of the servers are termed high or low relative to one another. There is a need to quantify the server capacity to overcome this subjective assessment. Subsequently, a method to split and distribute the requests based on this quantified server capacity is also needed.

The above issues have been addressed by many existing techniques and the exhaustive study of these literatures helped to formulate the objectives of the research. This research intends to devise a request scheduling principle for a heterogeneous distributed system having limited resources. The research work brings out,

i) A method to identify the most preferred common attribute of a set of servers, based on whose values the capacity of each server can be measured in a dissimilarly configured server pool.

ii) A method to determine each server’s allocation share.

iii) An optimal request scheduling technique that assigns the requests to servers.

iv) A customizable version of the identified solution to solve the request scheduling problem of any distributed system with suitable parameters.

v) A simulator implementing the designed scheduling principle.

Suitable experiments were conducted and the experimental results show considerable improvement in the performance of the designed request scheduling principle compared to a few other existing principles. Areas of further improvement have also been identified and presented.