Abstract

Samah A Senbel

Scalable and Self-Organizing Server Clusters for QoS-aware Applications

We present a new scalable and decentralized design for a server cluster that provides a single service with a certain QoS objective: maximum response time guarantee. Our design is based upon providing autonomy for each node in the cluster by having the node maintain only partial information about the system and having a mechanism for each node to process jobs redirect it to a neighbor node. Also, each node decides whether it should remain an active node in the system got to a standby mode, based upon information about its neighbor node and its own utilization. The metrics for the system performance are the rate of rejected requests in the system, the average experienced response time for service requests, the system size, and the ability of the system to react rapidly to changes in external load node failures. Our system performed well in all metrics compared to a similar system that provides the same goals