Abstract

Samah A Senbel

Pro-active task sharing in a self-organizing server cluster

We present a dynamic self-organizing design for a server cluster that dynamically allocates jobs to servers in a server cluster, while maintaining certain QoS objectives, mainly a guaranteed maximum response time. 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. We introduce a new technique for load balancing to be implemented by nodes with low utilization that drastically reduces the job rejection ratio as well as the average response time. The basic idea is for low-utilization nodes to aggressively advertise themselves to high-utilization nodes as candidate neighbors, and to attempt to grab jobs from the job queues of more busy neighbors. Our new technique is compared to other techniques and it performed well in all metrics: rejection ratio, average job response time, average node utilization, and reaction to an overload situation.