Abstract

Samah A Senbel

Load-balancing in a self-organizing server cluster using local leaders

This paper presents a dynamic self-organizing design for a server cluster that supports multiple services with different priorities. Our system dynamically allocates jobs to servers in a server cluster, while maintaining certain QoS objectives for each service. A "Local leader" is dynamically set up for each service, and it controls the membership of its service, and thus the cluster size. It communicates with other local leaders to maintain the system balance. We introduce a new technique for load balancing within a single service 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. Load balancing between services is performed between local leaders to optimize the overall performance. The proposed technique performs well in all performance metrics: rejection ratio, average job response time, average node utilization, Server assignment, and reaction to an overload situation.