Abstract

A server cluster consists of multiple simultaneously operating servers. Demands arrive to this cluster, which are distributed by a dispatcher between the queues of the different servers by some load balancing principle. The demands wait in these queues until they are served by the servers. These clusters offer an excellent way to model real life systems, like computer servers, the cash registers of shops or transportation networks.

When we examine the performance of these systems, the load balancing principle is a quite important parameter. We will examine five different principles in this paper. The Random assignment principle is the most trivial of them, it simply means that the arriving demands join a queue at random (uniformly). The Join-Idle-Queue (JIQ) principle means that the arriving demand joins an empty queue, if there is one, and otherwise it joins a queue randomly. The Join-Shortest-Queue principle is the most efficient of the discussed ones, it means that the arriving demand joins one of the queues with the fewest demands. A modified version of the previous is the Join-Shortest-Queue(d) (JSQ(d)) principle, which means that the dispatcher chooses d queues at random, and sends the demand to the shortest one. The last principle we discuss is Join-Below-Threshold (JBT), it means that a threshold is assigned to the servers (it may differ between the different servers), and if there are servers with fewer demands than the threshold, then the demand is sent to one of these, and otherwise it is sent randomly.

In this paper we examine a simpler mathematical model of these clusters. We ignore some problems real clusters face, like the cost of communication between the servers and the dispatcher. We also consider the system to be a continuous time Markov chain (where the state space is given by the queue lengths of the servers), and for that we need that the arrival and the service of the demands are both Poisson point processes.

It is possible that inside the cluster all the servers have the same properties (in this case we talk about a homogeneous cluster), or there can be different server types as well (in this case the cluster is heterogeneous). The speed of the servers is determined by their service rates, which can be constant, or it can depend on the length of the queues, and in the case of heterogeneous clusters they can differ between the server types.

The state of the cluster converges to a given set of functions (in which the functions represent the number of the servers of a given queue length in a server type in respect to the number of all servers in the cluster, depending on the elapsed time). These functions also converge to a stationary state, as the elapsed time goes to infinity.

In this paper we examine different clusters, especially focusing on their stationary states, simulating them and also numerically solving the differential equations which determine the functions. We will also calculate the mean and distribution of the system time of said clusters, further assisting the comparison of these different load balancing principles.