A System for Dynamic Server Allocation in Application Server Clusters

A.P. Chester

Abstract

Application server clusters are often used to service high-throughput web applications. In order to host more than a single application, an organization will usually procure a separate cluster for each application. Over time the utilization of the two clusters will vary, leading to variation in the response times experienced by users of the applications. The environment considered here is shown in figure 1

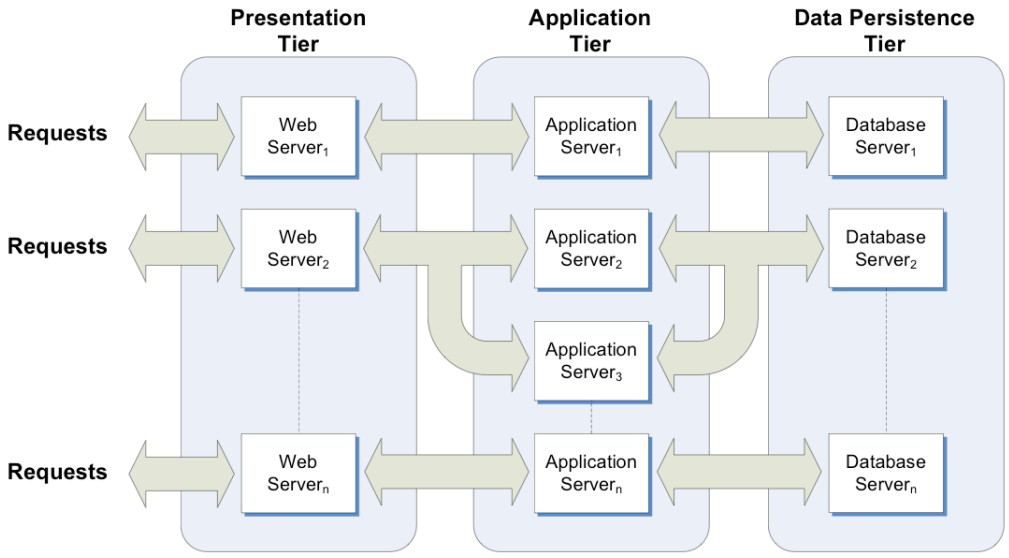

Figure 1: Multiple application architecture.

Techniques that statically assign servers to each application prevent the system from adapting to changes in the workload, and are thus susceptible to providing unacceptable levels of service. This paper investigates a system for allocating server resources to applications dynamically, thus allowing applications to automatically adapt to variable workloads. Such a scheme requires meticulous system monitoring, a method for switching application servers between *server pools* and a means of calculating when a server switch should be made (balancing switching cost against perceived benefits).

Experimentation is performed using such a switching system on a Web application testbed hosting two applications across eight application servers. The test bed is used to compare several theoretically derived switching policies. The Average Flow switching policy is shown to provide the best policy, when considering the mean response times for this application.