Abstract

E-commerce is growing to represent an increasing share of overall sales revenue, and online sales are expected to continue growing for the foreseeable future. This growth translates into increased activity on the supporting infrastructure, leading to a corresponding need to scale the infrastructure. This is difficult in an era of shrinking budgets and increasing functional requirements. Increasingly, IT managers are turning to virtualized cloud providers, drawn by the pay-for-use business model. As cloud computing becomes more popular, it is important for data center managers to accomplish more with fewer dollars (i.e., to increase the utilization of existing resources). Advanced request distribution techniques can help ensure both high utilization and smart request distribution, where requests are sent to the service resources best able to service them. While such request distribution techniques have been applied to the web and application layers of the traditional online application architecture, request distribution techniques for the data layer have focused primarily on online transaction processing scenarios. However, online applications often have a significant read-intensive workload, where read operations constitute a significant percentage of workloads (up to 95 percent or higher). In this paper, we propose a cost-based database request distribution (C-DBRD) strategy, a policy to distribute requests, across a cluster of commercial off-the-shelf databases, and discuss its implementation. We first develop the intuition behind our approach, and describe a high-level architecture for database request distribution. We then develop a theoretical model for database load computation, which we use to design a method for database request distribution and build a software implementation. Finally, following a design science methodology, we evaluate our artifacts through experimental evaluation. Our experiments, in the lab and in production-scale systems, show significant improvement of database layer resource utilization, demonstrating up to a 45 percent improvement over existing request distribution techniques.

Keywords: Database clusters, request distribution, task allocation, design research