Abstract

When deploying networked computing-based applications, proper resource management of the server-side resources is essential for maintaining quality of service and cost efficiency. The work presented in this thesis is based on six papers, all investigating problems that relate to resource management of server-based systems. Using a queueing system approach we model the performance of a database system being subjected to write-heavy traffic. We then evaluate the model using simulations and validate that it accurately mimics the behavior of a real test bed. In collaboration with Ericsson we model and design a per-request admission control scheme for a Mobile Service Support System (MSS). The model is then validated and the control scheme is evaluated in a test bed. Also, we investigate the feasibility to estimate the state of a server in an MSS using an event-based Extended Kalman Filter. In the brownout paradigm of server resource management, the amount of work required to serve a client is adjusted to compensate for temporary resource shortages. In this thesis we investigate how to perform load balancing over self-adaptive server instances. The load balancing schemes are evaluated in both simulations and test bed experiments. Further, we investigate how to employ delay-compensated feedback control to automatically adjust the amount of resources to deploy to a cloud application in the presence of a large, stochastic delay. The delay-compensated control scheme is evaluated in simulations and the conclusion is that it can be made fast and responsive compared to an industry-standard solution.