Abstract

This thesis addresses challenges in increasing the robustness of cloud-deployed applications and services to unexpected events and dynamic workloads. Without precautions, hardware failures and unpredictable large traffic variations can quickly degrade the performance of an application due to mismatch between provisioned resources and capacity needs. Similarly, disasters, such as power outages and fire, are unexpected events on larger scale that threatens the integrity of the underlying infrastructure on which an application is deployed.

First, the self-adaptive software concept of brownout is extended to replicated cloud applications. By monitoring the performance of each application replica, brownout is able to counteract temporary overload situations by reducing the computational complexity of jobs entering the system. To avoid existing load balancers interfering with the brownout functionality, brownout-aware load balancers are introduced. Simulation experiments show that the proposed load balancers outperform existing load balancers in providing a high quality of service to as many end users as possible. Experiments in a testbed environment further show how a replicated brownout-enabled application is able to maintain high performance during overloads as compared to its non-brownout equivalent.

Next, a feedback controller for cloud autoscaling is introduced. Using a novel way of modeling the dynamics of typical cloud application, a mechanism similar to the classical Smith predictor to compensate for delays in reconfiguring resource provisioning is presented. Simulation experiments show that the feedback controller is able to achieve faster control of the response times of a cloud application as compared to a threshold-based controller.

Finally, a solution for handling the trade-off between performance and disaster tolerance for geo-replicated cloud applications is introduced. An automated mechanism for differentiating application traffic and replication traffic, and dynamically managing their bandwidth allocations using anMPC controller is presented and evaluated in simulation. Comparisons with commonly used static approaches reveal that the proposed solution in overload situations provides increased flexibility in managing the trade-off between performance and data consistency.