© 2013 FIZ Karlsruhe

io-port 06055971

Muppala, Sireesha; Zhou, Xiaobo; Zhang, Liqiang; Chen, Guihai

Regression-based resource provisioning for session slowdown guarantee in multi-tier Internet servers.

J. Parallel Distrib. Comput. 72, No. 3, 362-375 (2012).

Summary: Autonomous management of a multi-tier Internet service involves two critical and challenging tasks, one understanding its dynamic behaviors when subjected to dynamic workloads and second adaptive management of its resources to achieve performance guarantees. We propose a statistical machine learning based approach to achieve session slowdown guarantees of a multi-tier Internet service. Session slowdown is the relative ratio of a session's total queueing delay to its total processing time. It is a compelling performance metric of session-based online transactions because it directly measures user-perceived relative performance and it is independent of the session length. However, there is no analytical model for session slowdown on multi-tier servers. We first conduct training to learn the statistical regression models that quantitatively capture an Internet service's dynamic behaviors as relationships between various service parameters. Then, we propose a dynamic resource provisioning approach that utilizes the learned regression models to efficiently achieve session slowdown guarantee under dynamic workloads. The approach is based on the combination of offline training and online monitoring of the Internet service behavior. Simulations using the industry standard TPC-W benchmark demonstrate the effectiveness and efficiency of the regression based resource provisioning approach for session slowdown oriented performance guarantee of a multi-tier e-commerce application.

Keywords: scalable multi-tier Internet services; autonomous resource provisioning; performance guarantee; session slowdown; regression modeling and analysis
doi:10.1016/j.jpdc.2011.11.011