A QoS based framework for efficient web servers

Nermin Mohamed Fawzy Mahmoud Salem

Abstract

Abstract - The explosive growth in popularity of the World Wide Web is leading to a number of performance problems such as decreased QoS (Quality of Service) when measured in terms of completed sessions and perceived service time by users in case of overloaded servers.

This paper contributes to research aiming to improve the QoS of commercial web servers. The QoS is defined in terms of server throughput and service time. The work is a two-sided approach for the problem. The first side is concerned with load balancing of web traffic, while the second tackles admission control for the server itself. The overall approach combines a locality aware solution (distributed workload request distribution policy) for the first problem with a non locality aware solution (predictive admission control) under CODA (Completely Distributed Architecture) for the second problem.
The solution methodology for the first part allows all server nodes to participate in session dispatching and exploit at maximum the good features of existing centralized algorithms. That of the second part uses an adaptive time slot scheduling that sets the execution times of the algorithm depending on the burstiness arriving to the system, with the aim of reducing the computational cost of the algorithm. Using a simulation model, we could show that a web server augmented with traffic control and session-based admission control performs better in terms of QoS than a server without these controls. This means it offers a better throughput and is able to provide a fair guarantee of completion, for any accepted session, independent of the session length.

Ain Shams Engineering Journal 2012, January