ELSEVIER

Performance Evaluation 42 (2000) 223–239

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Business-oriented resource management policies for e-commerce servers☆

Daniel A. Menascé[a,*], Virgilio A.F. Almeida[b], Rodrigo Fonseca[b], Marco A. Mendes[b]

[a] *Department of Computer Science, MS4A5, George Mason University, Fairfax, VA 22030, USA*
[b] *Department of Computer Science, University of Federal de Minas Gerais, Belo Horizonte, MG 30161, Brazil*

Abstract

Quality of service of e-commerce sites has been usually managed by the allocation of resources such as processors, disks, and network bandwidth, and by tracking conventional performance metrics such as response time, throughput, and availability. However, the metrics that are of utmost importance to the management and shareholders of a Web store are revenue and profit. Thus, resource management schemes for e-commerce servers should be geared towards optimizing business metrics as opposed to conventional performance metrics. This paper uses a state transition graph called customer behavior model graph (CBMG) to describe a customer session. It then presents a family of priority based resource management policies for e-commerce servers. Priorities change dynamically as a function of the state a customer is in and as a function of the amount of money the customer has accumulated in his/her shopping cart. A detailed simulation model was developed to assess the gain of these dynamic policies with respect to policies that are oblivious to economic considerations. Simulation results show that the multilevel dynamic priority scheme suggested here can significantly improve the values of business-oriented metrics, such as revenue per second, during peak periods. E-commerce sites that use this approach will be able to improve revenue at peak times with the same server capacity.

Keywords: E-commerce; Resource management; Customer behavior model graph; Workload characterization; Business-oriented metrics

1. Introduction

It has been recognized by many that congestion and poor performance can be the major impediments for the success of e-commerce. Many e-commerce sites, especially those in the financial trading business, have been facing serious problems and financial losses when customers are not allowed to trade in a timely manner. Some disgruntled customers sue on-line trading services if they feel that they have been short changed and others just move their business elsewhere.

IT managers of Web stores have been managing allocation of resources such as processors, disks, and networks, by tracking conventional performance metrics such as response time, throughput, and

☆ The subject matter of this paper is covered by a pending patent application at the United States Patent and Trademark Office.

* Corresponding author.

E-mail addresses: menasce@cs.gmu.edu (D.A. Menascé), virgilio@dcc.ufmg.br (V.A.F. Almeida), rfonseca@dcc.ufmg.br (R. Fonseca), corelio@dcc.ufmg.br (M.A. Mendes).

0166-5316/00/$ – see front matter
PII: S0166-5316(00)00034-1

availability. However, the metrics that are of utmost importance to the management of a Web store are revenue and profit. Thus, resource management schemes for e-commerce servers should be aimed at optimizing business-oriented metrics such as revenue per second instead of focusing on conventional performance metrics.

Resource management policies for e-commerce sites should be based on the behavior of customers and on how they change state as they navigate through the site, going from browsing to searching, selecting items, adding them to their shopping carts and paying. We present in this paper a family of priority based resource management policies for e-commerce servers. Priorities change dynamically as a function of the state a customer is in, as a function of the user profile, and as a function of the amount of money the customer has accumulated in his/her shopping cart. The policies can also be tuned to provide good performance to customers who are just entering the Web store even before they add any items to their shopping carts. We believe that resource management policies such as the ones presented in this paper should be integrated into future commercial e-commerce products. This would allow e-commerce servers to handle peak loads with existing resources in a way that minimizes revenue loss due to poor quality of service.

A detailed simulation model was developed to assess the gain of our policies with respect to policies that are oblivious to monetary considerations. The simulation uses the same probability distributions employed by SURGE [2], a workload generator for Web sites, augmented by a generator of e-commerce requests that mimic typical customer behavior. Requests generated by a customer within a session are generated from a customer behavior model graph (CBMG) that captures how users navigate through the site. The CBMG representation is used in this paper as a means of characterizing workloads for e-commerce sites. As an example, two types of customer profiles, heavy and occasional, were considered. The results of our simulations show that the dynamic priority scheme introduced in this paper can increase business-oriented metrics such as revenue per second by almost 30% over the no priority case.

The rest of this paper is organized as follows. Section 2 discusses new metrics for e-commerce sites. Section 3 describes e-commerce workloads as composed of session requests and CBMGs. Section 4 describes new resource management policies for e-commerce servers. Section 5 describes the simulator and the simulation environment used to analyze the new policies proposed. Section 6 describes the numerical results obtained. Section 7 compares our work with that of others. Finally, Section 8 presents some concluding remarks.

2. Novel metrics for e-commerce

The quality of the service provided by online information systems, such as Web servers, has been traditionally assessed by metrics such as response time, throughput, reliability, and availability. Response time can be measured at the server side, in which case it does not include any client and external network time, or it can be measured from the user's perspective, in which case it includes components such as browser time, network access time at the client side, ISP time (at both ends), Internet time, network access time at the server side, and server response time.

Throughput is usually measured in requests per second or transactions per second and determines the rate at which the system can deliver work. While throughput is important from the perspective of the site administrator, it is irrelevant to end-users who are concerned about the response time they get. When the response time is too high, users complain if they have no choice but use the system. In e-commerce, customers usually have a choice: they leave the site and move to another Web store. This translates into lost revenue for the Web store and decreased throughput.