Performance and Availability Analysis of an E-Commerce Site

Swapna S. Gokhale and Jijun Lu
Department of Computer Science and Engineering
University of Connecticut, Storrs, CT 06269, USA
{ssg, jijun.lu}@engr.uconn.edu

Abstract

To encourage customers to prefer e-commerce services over their on-site counterparts, it is necessary that e-commerce services be offered with superior performance and availability. A systematic methodology to enable a quantitative analysis of the performance and availability attributes of an e-commerce site can assist in identifying bottlenecks and suggesting strategies for improvement. The development of such a methodology is the subject of this paper. The analysis methodology derives expressions for performance and availability metrics of a site which are important from a customer's perspective, namely, the session length and session availability. The methodology is hierarchical, in the first step expressions are derived for each user group based on the navigational pattern of the group. The navigational pattern of a user group is represented by a Customer Behavior Model Graph (CBMG), which is mapped to a discrete time Markov chain (DTMC) for analysis. In the second step, expressions for the session length and availability of the site are derived using the distribution of the customer groups and the expressions for each customer group. We illustrate the potential of the methodology with a case study.

1 Introduction and motivation

E-commerce, which involves commercial transaction of goods and services between different parties electronically mainly through open Internet-based systems rather than physical exchange or contact is rapidly becoming a prominent way of conducting business. A number of companies and stores already provide e-commerce services for Internet users in addition to the traditional on-site services and this number is expected to rise dramatically as the use of the Internet becomes prevalent and companies discover and appreciate the cost effectiveness of doing business over the web. To encourage customers to prefer e-commerce services over their traditional on-site counterparts, it is necessary that an e-commerce site offer online services with superior performance and availability. Furthermore, since e-commerce sites are typically involved in transaction-based financial services such as online payment, banking and stock trading, discontinuity of service is unacceptable from the customers' point of view [2, 15, 18]. A customer unsatisfied with the quality of the services is likely to cancel the transaction and may even turn to the competitor's sites for similar products/services. At the minimum, this will result in a financial loss for the service provider. Moreover, this could also tarnish the reputation of the service provider and may discourage the customer from ever using the site again. Thus, performance and availability is of paramount importance in ensuring the success of an e-commerce site. A systematic, quantitative performance and availability analysis methodology will assist in identifying bottlenecks and suggesting ways of improving these attributes.

A customer[1] of an e-commerce site interacts with the site through a series of consecutive and related requests. This sequence of related requests is termed as a session and the request sequence in a single session is known as the navigational pattern. Session performance and availability of a user will depend on the user's navigational pattern and the performance and the availability of the site for each request types for that user and these factors must be considered in the analysis methodology. Further, it must be possible to obtain aggregate performance and availability metrics for the site as a whole across all user groups, based on the distribution or mix of the user groups.

In this paper we describe a performance and availability analysis methodology for an e-commerce site. We consider the metrics that are important from a customer's perspective, namely, the session length and session availability. The analysis methodology is hierarchical, in that in the first step it derives analytical expressions for the session length and session availability for a single group of customers, considering the navigational pattern of the group and the performance and availability metrics of the site for that group. The navigational pattern of a group of cus-

[1] The terms user and customer are used interchangeably in this paper.

Proceedings of the 30th Annual International Computer Software and Applications Conference (COMPSAC'06)
0-7695-2655-1/06 $20.00 © 2006 IEEE

tomers is represented by a Customer Behavior Model Graph (CBMG) [15]. The CBMG is mapped to a discrete time Markov chain (DTMC) [20] to facilitate the analysis. In the second step, analytical expressions for the overall session length and availability of the site, based on the distribution of the customer groups and the performance and availability expressions for each group are derived. We illustrate the potential of the methodology using a case study.

The layout of the paper is as follows: Section 2 provides an overview of DTMCs, which is the modeling paradigm used in the performance and availability analysis methodology. Section 3 provides an overview of CBMG used to represent the navigational pattern of a single user group. It also describes the process of mapping a CBMG to a DTMC for analysis. Section 4 presents the hierarchical performance and availability analysis methodology. Section 5 illustrates the potential of the methodology with a case study. Section 6 summarizes the related research and places our work in the context of the prevalent efforts. Section 7 offers concluding remarks and directions for future research.

2 Overview of DTMCs

In this section we provide an overview of discrete time Markov chains (DTMCs) which is the modeling paradigm used to represent a CBMG for the purpose of performance and availability analysis. A CBMG will represent the navigational pattern that would be observed over a single session, with a definite starting point and a definite ending point. Since these characteristics fit the description of an absorbing DTMC, we elaborate on absorbing DTMCs in this section. An interested reader may refer to [20] for details.

A DTMC is characterized by its one-step transition probability matrix $\mathbf{P} = [p_{i,j}]$. $\mathbf{P}$ is a stochastic matrix since all the elements in a row of $\mathbf{P}$ sum to 1.0, and each element lies in the range $[0, 1]$. A DTMC is said to be absorbing if there is at least one state from which there is no outgoing transition. A DTMC upon reaching the absorbing state is destined to remain there forever. The transition probability matrix of an absorbing DTMC can be partitioned as:

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{C} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad (1)$$

where $\mathbf{Q}$ is a $(n-m) \times (n-m)$ substochastic matrix (with at least one row sum less than 1), $\mathbf{I}$ is an $m \times m$ identity matrix, $\mathbf{0}$ is a $m \times (n-m)$ matrix of zeros, and $\mathbf{C}$ is a $(n-m) \times m$ matrix, with m absorbing states in a chain of n states.

The k-step transition probability matrix $\mathbf{P}^k$ has the form:

$$\mathbf{P}^k = \begin{pmatrix} \mathbf{Q}^k & \mathbf{C}' \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad (2)$$

where the entries of matrix $\mathbf{C}'$ are not relevant. The (i,j)-th entry of matrix $\mathbf{Q}^k$ denotes the probability of arriving in transient state j, starting from transient state i in exactly k steps. It can be shown that $\sum_{k=0}^{t} \mathbf{Q}^k$ converges as t approaches infinity. This implies that the inverse matrix $(\mathbf{I} - \mathbf{Q})^{-1}$ called the fundamental matrix $\mathbf{M}$ exists and is given by:

$$\mathbf{M} = \mathbf{I} - \mathbf{Q}^{-1} = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \ldots = \sum_{l=0}^{\infty} \mathbf{Q}^l \quad (3)$$

Without loss of generality, we assume that the process begins in state 1 and let $V_{1,j}$ denote the number of visits to state j starting from state 1 before the process is absorbed. The $(1,j)$-th entry of the fundamental matrix $\mathbf{M}$, $m_{1,j}$, represents the expected number of visits to state j starting from state 1 before the process is absorbed. If ν_j denotes the expected value of $V_{1,j}$, then we have:

$$\nu_j = E[V_{1,j}] = m_{1,j} \quad (4)$$

The fundamental matrix can also be used to compute the variance of the number of visits to state j. Let ${\sigma_j}^2$ denote the variance of the number of visits to state j starting from state 1. Then, we have:

$$\boldsymbol{\sigma^2} = \mathbf{M} \times (2\mathbf{M_{dg}} - \mathbf{I}) - \mathbf{M_{sq}} \quad (5)$$

where $\mathbf{M_{dg}}$ represents a diagonal matrix and $\mathbf{M_{sq}}$ denotes a matrix where each entry in $\mathbf{M_{sq}}$ is the square of each entry in the fundamental matrix $\mathbf{M}$. Thus

$$Var[V_{1,j}] = \sigma_j^2 \quad (6)$$

3 E-commerce workload

This section provides an overview of Customer Behavior Model Graph (CBMG) [15] used to represent the navigational pattern of a group of users in a typical session. We also discuss how a CBMG can be mapped to a DTMC, which forms the basis of the performance and availability analysis methodology discussed in the subsequent section.

3.1 Overview of CBMG

In this section we provide an overview of CBMG, a detailed description is provided in [15]. A CBMG is a state transition graph proposed for the workload characterization of an e-commerce site. In a CBMG, states represent the type of requests that can be presented to an e-commerce site. Transitions between the states represent the sequencing among these request types. The transitions among the states is usually probabilistic, which captures the random nature of the sequence of requests that may be presented. In addition to the request types and probabilistic transitions among

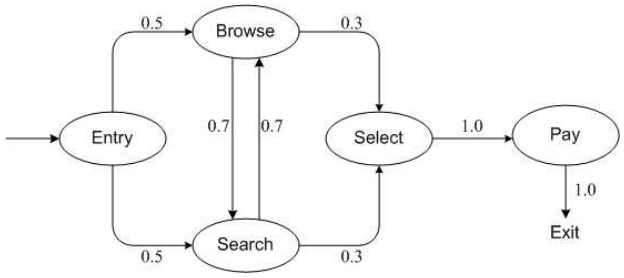

Figure 1. Example of a CBMG

the request types, CBMG also includes the time spent by a customer in thinking between two types of requests.

Formally, a CBMG can be represented by a pair $(\mathbf{W}, \mathbf{Z})$ of $n \times n$ matrices, where n is the number of states. $\mathbf{W} = [w_{i,j}]$ contains the transition probabilities between the states. The matrix $\mathbf{Z} = [z_{i,j}]$ contains the average think times between consecutive pairs of requests i and j. Figure 1 shows a simple CBMG of an e-commerce site with five types of requests that can be presented to the site.

A CBMG may be automatically extracted from web log files [3]. In the early stages, when the system is under development, the CBMG may also be extracted from the design documents. For instance, if the functional structure of a site is depicted using UML, the CBMG would be just a flat representation of the top level statechart describing the logic of the navigation possibilities. The CBMG can be used to calculate metrics such as the average number of requests of a single and all types and the resource usage of a session [3].

3.2 Mapping CBMG to DTMC

A CBMG is mapped to a DTMC according to the following process: Each state in a CBMG which represents a single type of service, is mapped to a state in the DTMC. Since these DTMC states represent request service by the site, we refer to these as service states. In order to account for think time between pairs of requests, additional states are introduced in the DTMC as follows. If a transition from node i to node j is allowed in the original CBMG (that is $w_{i,j} > 0$), indicating that there is a non-zero probability that a user may initiate a type j request after a type i request is completed, then a state $th_{i,j}$ is added to the DTMC. The states $th_{i,j}$ represent the think times between consecutive requests of type i and j and are referred to as think states. The transition probabilities of the DTMC with the additional states representing think times are obtained from the transition probabilities among the nodes of the CBMG as follows. If $w_{i,j} > 0.0$, then the transition probability between states $i, th_{i,j}$ is set to $w_{i,j}$. The transition probability among states $th_{i,j}, j$ is set to 1.0.

A CBMG is defined for each customer group that is served by the site. For example, TPC-W defines a CBMG for three groups of customers [19]. For each group, the DTMC generated by mapping the CBMG can be solved to obtain the visit statistics which consist of the mean and the variance of the number of visits to service states i and think states $th_{i,j}$. The visit statistics to state i indicate the mean and the variance of the number of times request type i is presented to the site. The visit statistics to state $th_{i,j}$ indicate the mean and the variance of the number of times request types i and j are presented consecutively to the site. We note that although the mean number of requests of each type can be obtained directly from a CBMG [3], mapping the CBMG to a DTMC allows us to compute additional statistics including the variance of the number of requests and the mean and the variance of the number of request pairs. Incorporating the variance effects can improve the availability estimates [5]. Further, the mean and the variance of request pairs allows us to incorporate think times to obtain a more accurate estimate of the session length.

4 Analysis methodology

The performance and availability analysis methodology is described in this section. The analysis methodology is hierarchical, in that in the first step, analytical functions which express session length and session availability in terms of the service and think state statistics, the service times and the availability metrics of individual request types, and think times and availability during think times are derived for a single group of customers. In the second step, the session lengths and availabilities for different groups of customers derived in the first step are composed to obtain aggregate metrics for the entire site. We let n denote the number of request types that can be presented to the site.

4.1 Customer group

In this section we derive analytical expressions for the session length and availability for a single customer group.

4.1.1 Performance analysis

The objective of performance analysis is to obtain an analytical expression for the mean or the expected session length in terms of the average service time of each request type and the average think time between the service request types for a single group of customers. For a customer group k, we let $z_{i,k}$ denote the average time taken by the e-commerce site to serve a request of type i and $z_{i,j,k}$ denote the average think time of the customer between successive requests of type i and type j. An important advantage of our methodology is that it can consider different service and think times for each group of customers. This is necessary, since the

service times of a site may differ based on the customer group. For example, in an online book store, customized personal pages may perform much more slowly for heavy buyers than for occasional buyers. This is because customized pages are created based on the previous orders and customer behavior and heavy buyers naturally have a larger volume of history that needs to be processed. Similar variations in the service and think times may be observed on auction sites where heavy users may have alerts on many auctions and also on customized news sites where frequent users may be subscribed to many new feeds than the ones who visit infrequently. Also, think times among request types may differ based on the request types. Referring to Figure 1, the think times between states $Browse$ and $Select$ ($Search$ and $Select$) may be higher than the think times between the states $Browse$ and $Search$, since the user may think longer to determine if he/she really wants to place an order.

The total time taken to serve all the requests of type i, denoted $T_{i,k}$ is given by Equation (7). Similarly, the total think time between requests of type i and type j, denoted $T_{i,j,k}$ is given by Equation (8).

$$T_{i,k} = X_{i,k} z_{i,k} \tag{7}$$

$$T_{i,j,k} = X_{th_{i,j,k}} z_{i,j,k} \tag{8}$$

In Equation (7), $X_{i,k}$ is the number of service requests of type i for a customer group k. Similarly, in Equation (8), $X_{th_{i,j,k}}$ is the number of times requests of type i and j are issued consecutively. We note that since $X_{i,k}$ and $X_{th_{i,j,k}}$ are random variables, $T_{i,k}$ and $T_{i,j,k}$ are also random variables. Based on the Taylor series expression for the mean of a function of a random variable [1], the expected total time taken to serve requests of type i in a typical session, denoted $E[T_{i,k}]$ is given by Equation (9). Similarly, the expected total think time between requests of type i and j denoted $E[T_{i,j,k}]$ is given by Equation (10).

$$E[T_{i,k}] = E[X_{i,k}] z_{i,k} = \nu_{i,k} z_{i,k} \tag{9}$$

$$E[T_{i,j,k}] = E[X_{th_{i,j,k}}] z_{i,j,k} = \nu_{th_{i,j,k}} z_{i,j,k} \tag{10}$$

In Equation (9), $\nu_{i,k}$ is the expected number of times requests of type i is presented in a given session by a group k customer. In Equation (10), $\nu_{th_{i,j,k}}$ is the expected number of times requests of type i and j are issued consecutively.

The expected session length for a customer of group k, denoted $E[T_k]$ is given by the sum of the expected total time taken to service all the requests and the expected total time spent in thinking between requests. From Equations (9) and (10), the expected session length is given by:

$$E[T_k] = \sum_{i=1}^{n} \nu_{i,k} z_{i,k} + \sum_{i=1,j=1}^{i=n,j=n} \nu_{th_{i,j,k}} z_{i,j,k} \tag{11}$$

The expected total think time in a session for a group k customer, denoted $E[T_{th,k}]$ is given by:

$$E[T_{th,k}] = \sum_{i=1,j=1}^{i=n,j=n} \nu_{th_{i,j,k}} z_{i,j,k} \tag{12}$$

The expected total think time is beyond the control of an e-commerce provider, but it influences the active duration of a customer session. An open, but idle customer session may still hold resources, which will ultimately impact the number of customers that can be served simultaneously [12].

4.1.2 Availability analysis

The objective of availability analysis is to obtain an expression for session availability in terms of the request and think state availabilities for a single group of customers. We let $a_{i,k}$ denote the site availability while serving a request of type i. Also, we let $a_{i,j,k}$ denote the availability when the customer is thinking between requests of type i and j.

The site availability while serving all requests of type i in a session, denoted $A_{i,k}$ is given by:

$$A_{i,k} = a_{i,k}^{X_{i,k}} \tag{13}$$

In Equation (13), $X_{i,k}$ is the number of service requests of type i in a session. Note that since $X_{i,k}$ is a random variable, $A_{i,k}$ is also a random variable. Using Taylor series expression for the mean of a function of a random variable [1], the expected site availability of type i requests in a given session, denoted $E[A_{i,k}]$, is given by:

$$E[A_{i,k}] = E[{a_{i,k}}^{X_{i,k}}] \approx a_{i,k}^{\nu_{i,k}} + \frac{1}{2} a_{i,k}^{\nu_{i,k}} (\log a_{i,k})^2 {\sigma_{i,k}}^2 \tag{14}$$

Using similar reasoning, the expected site availability in think state $th_{i,j}$, denoted $E[A_{i,j,k}]$ is given by:

$$E[A_{i,j,k}] =\approx a_{i,j,k}^{\nu_{i,j,k}} + \frac{1}{2} a_{i,j,k}^{\nu_{i,j,k}} (\log a_{i,j,k})^2 {\sigma_{i,j.k}}^2 \tag{15}$$

In Equations (14) and (15), $\nu_{i,k}$, ${\sigma_{i,k}}^2$, $\nu_{i,j,k}$, ${\sigma_{i,j,k}}^2$ have the same interpretation as in Section 4.1.1.

The session availability of the site for a customer belonging to group k, denoted A_k is given by the product of the expected site availabilities in all the service and think states:

$$A_k = \prod_{i=1}^{n} A_{i,k} \prod_{i=1,j=1}^{i=n,j=n} A_{i,j,k} \tag{16}$$

From Equations (14) and (15), the expected site availability of all request types, denoted $E[A_k]$, is given by:

$$\begin{aligned} E[A_k] \quad \approx \quad & \prod_{i=1}^{n} [a_{i,k}^{\nu_{i,k}} + \frac{1}{2} a_{i,k}^{\nu_{i,k}} (\log a_{i,k})^2 {\sigma_{i,k}}^2] \\ & \prod_{i=1}^{n} [a_{i,j,k}^{\nu_{i,j,k}} + \frac{1}{2} a_{i,j,k}^{\nu_{i,j,k}} (\log a_{i,j,k})^2 {\sigma_{i,j,k}}^2] \end{aligned} \tag{17}$$

4.2 Overall site

In this section we derive expressions for the aggregate session length and availability of the site across all customer groups. We let m denote the number of customer groups served by the site. We let o_i, $1 \leq i \leq m$ denote the probability that a customer belongs to group i.

The aggregate session length $E[T]$ of the site across all customer groups is given by Equation (18), where $E[T_k]$ is given by Equation (11).

$$E[T] = \sum_{k=1}^{m} E[T_k] o_k \tag{18}$$

Similarly, the aggregate session availability of the site, computed across all customer groups is given by Equation (19), where $E[A_k]$ is given by Equation (17).

$$E[A] = \sum_{k=1}^{m} E[A_k] o_k \tag{19}$$

5 Illustrations

We illustrate the analysis methodology described in Section 4 with a case study reported in [15].

System description: The e-commerce site under consideration is an online bookstore. This site provides online services such as user authentication, product searching, online payment, etc. A registered user starts a session by logging into this web site. The user will then browse for a book from the book list or perform a search for the book(s) he/she is interested in, by providing certain key word(s). A book is selected and added to the user's shopping cart before the user finally pays for it with different payment options. A heavy buyer is more likely to place an order or make a payment if he/she finds an interesting book, while an occasional buyer typically spends most of the time in browsing and searching without placing an order at the end of the session.

Figures 2 and 3 show the CBMGs for occasional and heavy buyers who conduct business with this e-commerce site [16]. There are seven states presented in each CBMG, i.e., Login, Browse, Search, Select, Add to Cart, Pay and Exit, which represent seven request types presented to the site. We assume that the customer mix consists of 80% occasional buyers and 20% heavy buyers.

Service and think state statistics: Figures 4 and 5 show the DTMCs for occasional and heavy buyers with states for think times. The solution of the DTMCs presented in these figures provide the service and think state statistics, which are summarized in Tables 1 and 2 respectively.

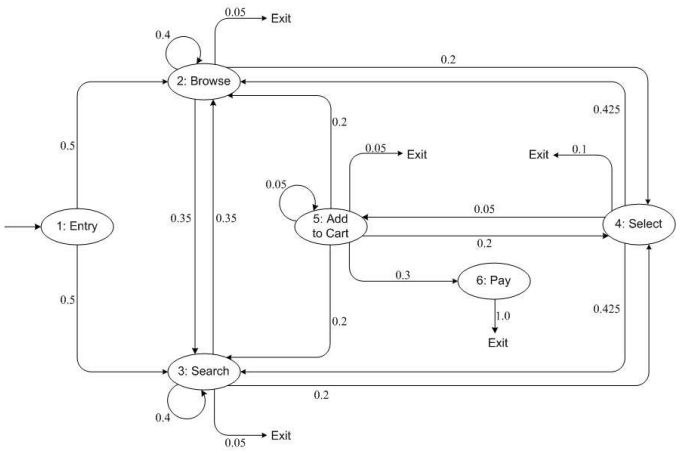

Figure 2. CBMG for an occasional buyer

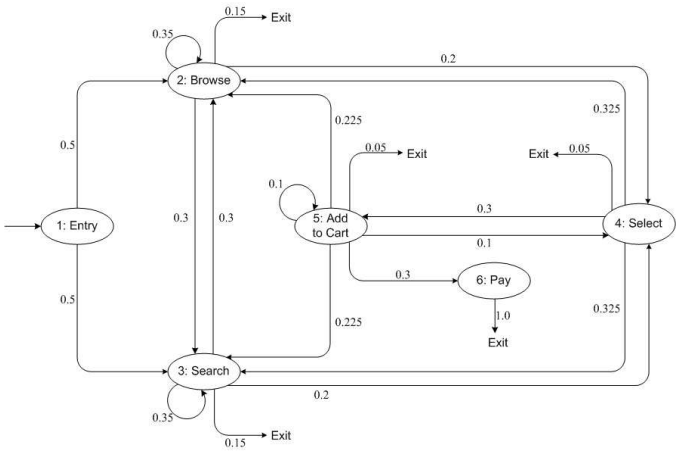

Figure 3. CBMG for a heavy buyer

Session length and availability: We now illustrate how the service and think state statistics computed from the DTMCs can be used for estimating the session length and session availability. The parameter settings in this section are used only for the sake of illustration. To use the methodology for the analysis of an e-commerce site, these parameters need to be obtained from different sources such as web logs, customer surveys, and expert opinion. In the early stages, when it is difficult to estimate these parameters, the analytical expressions in this paper can facilitate an analysis of the sensitivity of the performance and availability metrics to the variations in the parameter values. As an example, we set the request availability to 0.95, think state availability to 1.00, request service time to 1.00s and think time between requests to 5.00s. These parameter values are used for both groups of buyers. The session length and session availability computed using these values are reported in Table 3.

Table 2. Think state statistics

State	Occasional buyer		Heavy buyer		State	Occasional buyer		Heavy buyer	
	$v_{th_{i,j}}$	${\sigma_{th_{i,j}}}^2$	$v_{th_{i,j}}$	${\sigma_{th_{i,j}}}^2$		$v_{th_{i,j}}$	${\sigma_{th_{i,j}}}^2$	$v_{th_{i,j}}$	${\sigma_{th_{i,j}}}^2$
$th_{1,2}$	0.5000	0.2500	0.5000	0.2500	$th_{1,3}$	0.5000	0.2500	0.5000	0.2500
$th_{2,2}$	2.7050	11.1612	0.9486	2.1979	$th_{2,3}$	2.3669	7.0971	0.8131	1.2174
$th_{2,4}$	1.3525	2.7440	0.5421	0.7346	$th_{2,7}$	0.3381	0.2238	0.4065	0.2413
$th_{3,2}$	2.3669	7.0971	0.8131	1.2174	$th_{3,3}$	2.7050	11.1612	0.9486	2.1979
$th_{3,4}$	1.3525	2.7440	0.5421	0.7346	$th_{3,7}$	0.3381	0.2238	0.4065	0.2413
$th_{4,2}$	1.1619	2.5118	0.3645	0.4973	$th_{4,3}$	1.1619	2.5118	0.3645	0.4973
$th_{4,5}$	0.1367	0.1436	0.3364	0.3805	$th_{4,7}$	0.2734	0.1986	0.0561	0.0529
$th_{5,2}$	0.0288	0.0296	0.0841	0.0912	$th_{5,3}$	0.0288	0.0296	0.0841	0.0912
$th_{5,4}$	0.0288	0.0300	0.0374	0.0409	$th_{5,5}$	0.0072	0.0080	0.0374	0.0462
$th_{5,6}$	0.0432	0.0413	0.1121	0.0996	$th_{5,7}$	0.0072	0.0071	0.0187	0.0183
$th_{6,7}$	0.0432	0.0413	0.1121	0.0996					

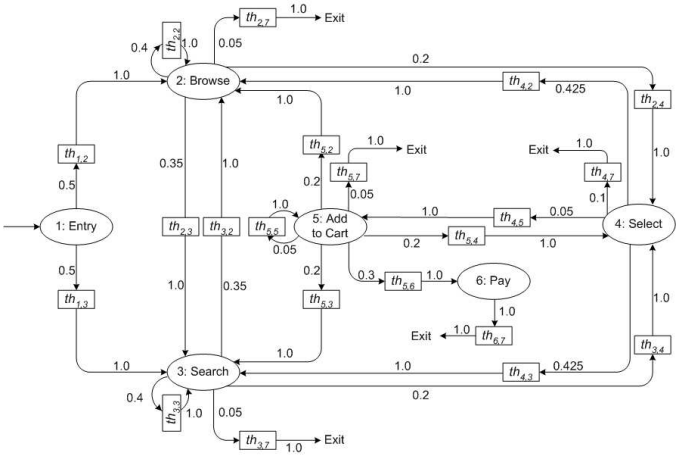

Figure 4. DTMC for an occasional buyer

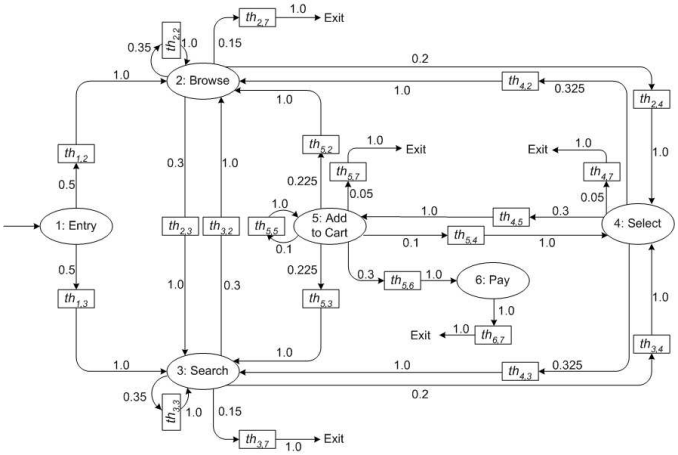

Figure 5. DTMC for a heavy buyer

Table 3. Session length and availability

Session metric	Occasional	Heavy	Aggregate
Availability	0.4417	0.6440	0.4822
Session length	105.68 s	49.17 s	94.38 s

Table 1. Service state statistics

State	Occasional buyer		Heavy buyer	
	v_i	${\sigma_i}^2$	v_i	${\sigma_i}^2$
Entry	1.000	0.0000	1.000	0.0000
Browse	6.7626	46.0885	2.7103	7.4883
Search	6.7626	46.0885	2.7103	7.4883
Select	2.7338	8.4768	1.1215	2.0229
Add	0.1439	0.1667	0.3738	0.5113
Pay	0.0432	0.0413	0.1121	0.0996
Exit	1.0000	0.0000	1.0000	0.0000

Discussion: The results in Table 3 indicate that the expected session length of an occasional buyer is longer than that of a heavy buyer. Intuitively, this is expected since a heavy buyer may typically visit a site with a preconceived idea of the books he/she would like to buy, due to which he/she may proceed to placing an order quickly. On the other hand, an occasional buyer may be mainly interested in browsing and searching and may buy something only if it really captures his/her interest. The expected session availability of a heavy buyer is higher than that of an occasional buyer. Once again, this is expected, since implicitly session availability is correlated with session length, for a longer

session is a result of a higher number of service requests presented to the site, which will result in lower availability.

The relationship between session length and availability and service times and availabilities of requests is complex and non-linear as exemplified by the equations derived in Section 4. In the present example, however, since the service times and availabilities for each request type are identical for both groups of customers, the service statistics can be used to provide guidance regarding how to improve site performance and availability. Referring to Table 1, it can be observed that the mean and the variance of the number of $Browse$ and $Search$ requests is very high for occasional buyers compared to heavy buyers. This indicates that not only do occasional buyers present a larger number of $Browse$ and $Search$ requests but they also exhibit a much greater variability in their navigational pattern. Thus, a cost-effective way of improving the session length and availability for occasional buyers is to improve the service times and availabilities of the $Browse$ and $Search$ requests. Indirectly, this would also improve the session length and availability of the heavy buyers, since the service statistics in Table 1 indicate that even for heavy buyers the $Browse$ and $Search$ requests dominate the navigational pattern. The difference in the service statistics between the $Browse$ and $Search$ requests and other request types is not as dramatic for heavy buyers as it is for occasional buyers. Thus, a given improvement in the service times and availabilities of $Browse$ and $Search$ requests would not result in as much improvement for heavy buyers as it would for occasional buyers.

6 Related research

In this section we provide a brief overview of the related research and place our work in context.

In the recent years, performance analysis of e-commerce sites has emerged as an active area of research. Gunther *et al.* [6] present a case study describing capacity planning of e-commerce sites. In their work, several spreadsheet-based techniques are used for forecasting both short-term and long-term trends in the consumption of server capacity. Two new performance metrics are introduced for site planning and procurement: effective demand, and doubling period. Hardwick *et al.* [7] use Indy, a new performance modeling framework, to create a performance analysis tool for database-backed web sites. This tool is validated using the predicted and observed performance of a sample e-commerce site. Kamra *et al.* [11] design a completely self-tuning admission controller for 3-tiered web sites based on classical control theoretic ideas, which is able to bound response times for requests while maintaining a high throughput under overload. Kohavi *et al.* [13] offer several recommendations for supplementary analysis which are very useful in practice. Menascé [14] reviews the important issues and challenges in modeling Web-based systems and discusses workload characterization, predictive models, and their use in various situations. Yang [8] present a performance study under the e-commerce workload with a mixture of static web page requests, CGI requests, servlet requests, and database queries. Iyengar *et al.* [9] present several techniques that can be used at popular sites to improve performance and availability based on two case studies. Datla *et al.* [4] conduct measurement-based performance analysis of an e-commerce application which uses Web services components. A session-oriented workload is generated according to a simplified TPC-W specification.

There also exist some efforts concerned with availability/dependability analysis of an e-commerce web site. Kaaniche *et al.* [10] present and illustrate a hierarchical modeling framework for the availability/dependability evaluation of an Internet-based travel agency example. The framework considers availability analysis at four different levels, namely, resource, service, function and user. Merzbacher *et al.* [17] present the results of a series of long-term experiments that measured the availability of selected web sites and services with the goal of duplicating the end-user experience. Based on these measurements, they propose a new metric for availability that goes beyond the traditional sole measure of uptime.

The methodology described in this paper is closest to function and service-level availability modeling considered in the hierarchical availability modeling framework presented by Kaaniche *et al.* [10]. Compared to their research however, the proposed methodology enjoys several advantages. First, their approach relies on enumerating paths through the graph representing the operational profile of the user, and path enumeration provides only an approximate availability estimate when the graph has loops. On the other hand, our approach maps the user CBMG to a DTMC, and hence we can consider the impact of loops analytically. Second, their approach is purely computational, which makes sensitivity and predictive analysis cumbersome. Using the analytical expressions produced by our methodology, sensitivity and predictive analysis and bottleneck identification is easily facilitated. Third, our methodology considers performance and availability in an integrated manner as opposed to theirs which focuses exclusively on availability.

7 Conclusions and future research

In this paper we presented a performance and availability analysis methodology for an e-commerce site. The methodology is hierarchical, in the first step analytical expressions for session length and availability in terms of the service and think time metrics are derived for a single group of customers, taking into consideration the navigational pattern of

the group. The navigational pattern of a group of users is represented using a CBMG, which is mapped to a DTMC for analysis. In the second step, expressions for aggregate session length and availability of the site based on the distribution of user groups are obtained. We illustrated the potential of the methodology using a case study.

Estimating the throughput of a site based on the idle time of each customer session is a topic of future research. Currently, service and think state availabilities are considered to be constant values. Incorporating time-dependent availability metrics is also a topic of future research. We also seek to develop techniques to extract CBMGs and service and think times and availabilities from different web artifacts such as transaction log files and design documents.

References

[1] L. J. Bain and M. Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, Belmont, CA, 1980.

[2] Y. Bakos. The emerging role of electronic marketplaces on the Internet. *Communications of the ACM*, 41(8):35–42, 1998.

[3] G. Ballocca, R. Politi, and G. Ruffo. Integrated techniques and tools for web mining, user profiling and benchmarking analysis. In *Proceeding of CMG'03*, 2003.

[4] V. Datla and K. Goseva-Popstojanova. "Measurement-based performance analysis of e-commerce applications with Web services components". In *IEEE International Conference on e-Business Engineering (ICEBE 05)*, pages 305–314, 2005.

[5] S. Gokhale. "Software reliability analysis incorporating second-order architectural statistics". *Intl. Journal of Reliability, Quality and Safety Engineering*, 12(3):267–290, 2005.

[6] N. J. Gunther. Performance and scalability models for a hypergrowth e-commerce web site. `http://arxiv.org/ftp/cs/papers/0012/0012022.pdf`.

[7] J. C. Hardwick, E., Papaefstathiou, and D. Guimbellot. Modeling the performance of e-commerce sites. In *Proceedings of the 27th International Conference of the Computer Measurement Group*, pages 3–12, 2001.

[8] X. He and Q. Yang. Performance evaluation of distributed web server architectures under e-commerce workloads. In *Proceedings of First International Conference on Internet Computing*, pages 285–292, 2000.

[9] A. Iyengar, J. Challenger, D. Dias, and P. Dantzig. High-performance web site design techniques. *IEEE Internet Computing*, 4(2):17–26, March 2000.

[10] M. Kaaniche, K. Kanoun, and M. Martinello. A user-perceived availability evaluation of a web based travel agency. In *Proceedings of the 2003 International Conference on Dependable Systems and Networks (DSN'03)*, pages 709–718, 2003.

[11] A. Kamra, V. Misra, and E. Nahum. Controlling the performance of 3-tiered web sites: modeling, design and implementation. In *SIGMETRICS 2004/PERFORMANCE 2004*, pages 414–415, 2004.

[12] M. Kihl, N. Widell, and C. Nyberg. Perfomance modelling and control of distributed E-commerce sites. In *Proceedings (on CD) of Infrastructure for e-Business, e-Education and e-Science on the Internet*, 2001.

[13] R. Kohavi and R. Parekh. Ten supplementary analyses to improve e-commerce web sites. In *Proceedings of the Fifth WEBKDD workshop (WEBKDD'03)*, pages 29–36, 2003.

[14] D. A. Menascé. Web performance modeling issues. *International Journal of High Performance Computing Applications*, 14(4):292–303, Winter 2000.

[15] D. A. Menascé and V. A. F. Almeida. *Capacity planning for web services: metrics, models and methods*. Prentice Hall, Upper Saddle River, NJ, 2002.

[16] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes. A methodology for workload characterization of E-Commerce sites. In *Proceedings of ACM Conference on E-Commerce*, pages 119–128, 1999.

[17] M. Merzbacher and D. Patterson. Measuring end-user availability on the web: Practical experience. In *Proceedings of the 2002 International Conference on Dependable Systems and Networks (DSN'02)*, pages 473–477, 2002.

[18] S. S. Y. Shim, V. S. Pendyala, M. Sundaram, and J. Z. Gao. Business-to-business e-commerce frameworks. *Computer*, 33(10):40–47, 2000.

[19] Transaction Processing Performance Council. TPC-W. `http://www.tpc.org/tpcw/`.

[20] K. S. Trivedi. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. John Wiley and Sons, 2001.