Modeling and Analysis of 2D Service Differentiation on e-Commerce Servers

Xiaobo Zhou, Jianbin Wei, and Cheng-Zhong Xu
Wayne State University, Detroit, Michigan 48202

***Abstract*— A scalable e-Commerce server should be able to provide different levels of quality of service (QoS) to different types of requests according to clients' navigation patterns and the server capacity. In this paper, we propose a two-dimensional (2D) service differentiation model for on-line transactions: *inter-session* and *intra-session*. The inter-session model aims to provide different levels of QoS to sessions from different customer classes, and the intra-session model aims to provide different levels of QoS to requests in different states of a session.**

A primary performance metric of on-line transactions is slowdown. It measures the waiting time of a request relative to its service time. We introduce service slowdown as a QoS metric of e-Commerce servers. It is defined as weighted sum of request slowdown in different sessions and different session states. We formulate the problem of 2D service differentiation provisioning as an optimization of processing rate allocation with the objective of minimizing service slowdown. We derive the optimal allocations for an M/G/1 server under various server load conditions and prove that the optimal allocations guarantees requests' slowdown to be square-root proportional to their pre-specified differentiation weights in both inter-session and intra-session dimensions. We evaluate the optimal allocation scheme via extensive simulations and compare it with another proportional slowdown differentiation scheme that is tailored from proportional queueing-delay differentiation in networking. Simulation results validate that both processing rate allocation schemes can achieve predictable, controllable, and fair 2D slowdown differentiation on e-Commerce servers. The optimal allocation scheme guarantees 2D service differentiation at a minimum cost of service slowdown.

I. Introduction

Service differentiation is to treat client requests differently based on clients' needs and servers' resource limitations. Because clients are different in their visiting patterns, receiving devices, and service fees, a scalable e-Commerce server needs to provide different levels of QoS to different clients. Service differentiation has been an active research topic in the arena of networking since its architecture was first formulated by IETF [6]. Its goal is to define configurable types of packet forwarding so as to provide per-hop differentiated services for large aggregate of network traffic. Network alone is not sufficient to support end-to-end service differentiation. There are recent efforts on server-side service differentiation for various Web and multimedia applications; see [2], [5], [7], [9], [12], [25] for representatives. However, few exists for service differentiation in session-based e-Commerce applications.

A session is a sequence of individual requests of different types made by a single customer during a single visit to an e-Commerce site. During a session, a customer can issue consecutive requests of various functions such as browse, search, select, add to shopping cart, register and pay. It has been observed that different customers exhibit different navigation patterns. Actually, only few customers are heavy buyers and all others are occasional buyers or visitors. Recent studies on customer behaviors of some e-Commerce sites showed that only 5% to 10% customers were interested in buying something during the current visit and about 50% of these customers were capable of completing their purchases [19], [20], [21]. Although it is important to accommodate the remaining 90% to 95% customers in order to turn them into loyal customers in future, the 5% to 10% premium customers should be preferential. This requires a scalable e-Commerce server to provide different levels of QoS to sessions from different customers. We refer to this as *inter-session service differentiation*.

An e-Commerce session contains a sequence of requests for various functions in different states. Requests in different states have different opportunities to turn themselves to be profitable. E-Commerce servers should also provide different levels of QoS to requests in different states in each session so that profitable requests like order and checkout are guaranteed to be completed in a timely manner. We refer to this as *intra-session service differentiation*.

In this paper, we investigate the problem of the two-dimensional (2D) service differentiation provisioning on e-Commerce servers, where the customers (and their requests) can be classified according to their profiles and shopping behaviors. User-perceived QoS of on-line transactions is often measured by a performance metric of response time. It refers to the duration of a request between its arrival and departure times, including waiting time in a backlogged queue and actual processing time. Network transmission delay is not considered because it is beyond the scope of this paper. Service differentiation provisioning with respect to response time can be achieved to some extent by conventional priority-based request scheduling. The principle of priority-based scheduling is widely used to support packet queueing-delay differentiation in networking. Most of the delay differentiation algorithms can be tailored for request response time differentiation on e-Commerce servers [8], [17].

Response time reflects user-perceived absolute performance of a server. It is not suitable for comparing the quality of requests that have different resource demands. Customers are likely to anticipate short delays for "small" requests like browsing, and are willing to tolerate long delays for "large" requests like search. A more important performance metric is *slowdown*. A request's slowdown is defined as the ratio of its delay in a backlogged queue relative to its service time. Since slowdown translates more directly to user-perceived system load, it is more often used as a performance metric of responsiveness on Interne servers [4], [13], [22], [25]. Service differentiation with respect to slowdown is beyond the capa-

bilities of priority based request scheduling schemes. Priority-based scheduling schemes adjust the priority of backlogged requests according to their experienced queueing delays, without taking into account any information about their service time. A queueing discipline based on request service time violates a fundamental Little's Law on which the priority-based scheduling principle is built.

In [25], Zhu, *et al.* developed a demand-driven node partitioning approach to provide different stretch factors (a variant metric of slowdown) to requests in different classes on clusters of servers. Their approach was targeted at stateless Web workload, instead of session-based e-Commerce transactions. This paper proposes a more general processing rate allocation strategy for 2D service differentiation provisioning with respect to slowdown. We assume that the e-Commerce server has a single processing resource bottleneck. Although processing a request often needs to consume resources of different types, resource management usually focuses on the allocation of the most critical resource. This single resource bottleneck assumption was made in [2], [9], [10], [13], [24], as well. The primary contributions of this paper are:

1) We propose a 2D service differentiation model for e-Commerce servers, namely, inter-session and intra-session service differentiation. We formulate the 2D service differentiation with respect to slowdown as an optimization problem of processing rate allocation for the objective of minimizing service slowdown of the server.
2) We derive an optimal processing rate allocation scheme and prove that the optimal allocation scheme guarantees square-root proportional service differentiation and hence it is fair.
3) We evaluate the allocation scheme via extensive simulations and verify the predictability, controllability, and fairness of the scheme. The simulation results show that the scheme can consistently achieve short-term and long-term 2D slowdown differentiation.

The rest of the article is organized as follows. Section II gives the 2D service differentiation model and a formulation of the processing rate allocation problem. Section III presents the optimal allocation scheme as well as a proportional slowdown differentiation scheme. Sections IV and V present implementation issues and simulation results. Related work is reviewed in Section VI. Section VII concludes the article.

II. 2D Service Differentiation

A. Modeling of 2D Service Differentiation

For service differentiation, incoming requests from different clients need to be classified into multiple classes according to their desired levels of QoS. The request classification can be done based on clients' profile, device, payment, etc. Basically, there are two types of service differentiation (DiffServ) schemes [6]. One is absolute DiffServ, in which each request class receives an absolute share of resource usages. The other is relative DiffServ, in which a class with a higher desired QoS level (referred to as higher class) will receive better (at least no worse) service quality than a lower class. Although absolute DiffServ is desired to Internet services like audio/video streaming applications that have hard time constraints, relative DiffServ is sufficient for soft real-time applications like e-Commerce transactions.

In order for a relative DiffServ scheme to be effective, the scheme must satisfy two basic properties: *predictability* and *controllability*. Predictability requires that higher classes receive better or no worse service quality than lower classes, independent of the class load distributions. Controllability requires that the scheduler contain a number of controllable parameters that are adjustable for the control of quality spacings between classes. An additional requirement on e-Commerce servers is *fairness*. That is, requests from lower classes should not be over-compromised for requests from higher classes. It is important to provide preferential treatments to sessions from premium customers and to requests that are likely to end with a purchase. However, e-Commerce servers should also handle other non-buying sessions that account for about 90% to 95% of visits if one wants to turn them into loyal customers [19], [20], [21].

We propose a 2D relative DiffServ model for inter-session and intra-session slowdown differentiation. Because different customers have different navigation patterns, the 2D service differentiation model classifies the customers into m classes according to statistics of their shopping behaviors, such as buy to visit ratio. Customers in the same class have similar navigation patterns. Actually, some e-Commerce sites request customers to login before they start to navigate through the site. Their profiles help simplify the customer classification. The 2D service differentiation model assumes that each session of customer class i ($1 \leq i \leq m$) has n states, each corresponding to a request function. A customer's request is classified as being in different states according to the type of function requested.

We assume that session arrivals from each customer class meet a Poisson process. We note that requests in each state from sessions of different customers are independent because the session head requests are independent. However, a customer may visit a state many times in a session. For example, a customer can submit a search request at time t_1, select a commodity at time t_2 (after some think time), and submit another search request at time t_3. Evidently, these requests at the search state are dependent and their dependency degree is determined by the navigation pattern of that customer. We notice that an e-Commerce server can accommodate many concurrent sessions from independent customers and that the number of re-visits in a session is limited (on average, the maximum number of visits at a state in a session is 2.71 and 6.76 for a heavy buyer class and for an occasional buyer class, respectively according to [19], [20]). That is, all the requests at the same state are weakly dependent. Therefore, we assume that request arrivals in each state from sessions of a customer class still meet a Poisson process. Our experimental results also verified this expectation. We model the server as

a M/G/1 queue. The requests are scheduled in a processor-sharing manner by storing them into $m \times n$ queues, each associated with a state.

Assume requests in Poisson process arrive at a rate λ. Denote μ the request processing rate. It follows that the traffic intensity $\rho = \lambda/\mu$. Let S be a request's slowdown. According to queueing theories [16], when $\rho < 1$ ($\lambda < \mu$), we have the expected slowdown as

$$S = \frac{\rho}{1-\rho}. \tag{1}$$

B. Formulation of Processing Rate Allocation

The basic idea of the processing rate allocation for provisioning 2D service differentiation on an e-Commerce server is to divide the scheduling process into a sequence of short intervals. In each interval, based on the measured resource utilization and the predicted workload, the available processing resource usages are allocated to requests in different states from different sessions.

Let C be the total amount of the processing resource available during the current resource allocation interval. The server's request scheduler has to determine the amount of the resource usages allocated to requests in each queue so that 2D service differentiation can be achieved and resource utilization can be maximized. Table I summarizes the notations used in the problem formulation.

TABLE I
NOTATION SUMMARY.

Symbol	Description
m	Number of customer classes
n	Number of states in a session
λ_i	Session arrival rate of customer class i
α_i	Quality weight of session i
β_j	Quality weight of requests in state j
C	Available processing resource
$c_{i,j}$	Allocated resource to requests in state j of session i
r_j	Average resource demand of a session in state j
$v_{i,j}$	Average number of visits to state j in session i
$d_{i,j}$	Demanded resource by requests in state j of session i
$s_{i,j}$	Slowdown of a request in state j of session i

A session in different states usually demands different processing resource usages [3], [19], [20], [23]. Let r_j be the average resource demand of a session in state j. Note that requests in the same state from different sessions generally demand the same amount of the processing resource. Let $c_{i,j}$ be the amount of the resource allocated to requests from sessions of class i ($1 \le i \le m$) in state j ($1 \le j \le n$) in the current allocation interval. Thus, $c_{i,j}/r_j$ is the processing rate of requests in state j from sessions of class i. Let $v_{i,j}$ denote the average number of visits to state j in a session from class i. Let $d_{i,j}$ denote the resource requirement of a session from class i in state j. That is, $d_{i,j} = v_{i,j} r_j$. According to (1), the slowdown of a request from a session of class i in state j, $s_{i,j}$, is calculated as:

$$s_{i,j} = \frac{\lambda_i v_{i,j}}{c_{i,j}/r_j - \lambda_i v_{i,j}} = \frac{\lambda_i d_{i,j}}{c_{i,j} - \lambda_i d_{i,j}}, \tag{2}$$

where λ_i is the session arrival rate of class i.

We consider the processing rate allocation for 2D service differentiation when the following constraint holds in each resource allocation interval:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_i d_{i,j} < C. \tag{3}$$

That is, the request processing rate of the server is higher than the request arrival rate. Otherwise, a request's slowdown can be infinite. Service differentiation provisioning would be infeasible.

Let α_i be the normalized quality differentiation weight of sessions from class i. That is, $\alpha_i > 0$ and $\sum_{i=1}^{m} \alpha_i = 1$. Because sessions from class i should receive better or no worse service quality than sessions from class $i+1$ according to inter-session service differentiation, without loss of generality, we assume $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_m$. The values of α_i can be determined according to the shopping behaviors of class i, such as their buy to visit ratio [19], [20].

Let β_j be the normalized quality differentiation weight of state j in a session. That is, $\beta_j > 0$ and $\sum_{j=1}^{n} \beta_j = 1$. The values of β_j can be determined according to the characterization of transition probability from state j to state pay in sessions from all classes. Without loss of generality, we assume $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_n$. Based on the concept of slowdown for individual requests, we define a metric of *session slowdown* as $\sum_{j=1}^{n} \beta_j s_{i,j}$ to reflect the weighted slowdown of requests in a session from class i. We further define a metric of *service slowdown* as $\sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i \beta_j s_{i,j}$ to reflect weighted session slowdown of sessions from all classes.

We then formulate the processing rate allocation problem for 2D service differentiation as the following optimization problem:

$$\text{Minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i \beta_j s_{i,j} \tag{4}$$

$$\text{Subject to} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} c_{i,j} \le C \tag{5}$$

$$s_{i,j} = \frac{\lambda_i d_{i,j}}{c_{i,j} - \lambda_i d_{i,j}} > 0. \tag{6}$$

The objective function (4) is to minimize the service slowdown of the server. It implies that sessions from higher classes get lower slowdown (higher QoS) and hence inter-session differentiation is achieved. It also implies that sessions in high states get lower slowdown (higher QoS) and hence intra-session differentiation is achieved. The rationale behind the objective function is its feasibility, differentiation predictability, controllability and fairness, as we discussed in Section II-A. (5) gives a resource allocation constraint over variables $c_{i,j}$. (6) ensures the positivity of slowdown. In the next section, we will show the optimal allocation scheme derived from the optimization model can meet those requirements.

III. Processing Rate Allocation Schemes

A. An Optimal Allocation Scheme

The above optimization problem is essentially a continuous convex separable resource allocation problem. According to theories of general resource allocation problems [14], its optimal solution occurs when the first order derivatives of the objective function (4) over variables $c_{i,j}, 1 \leq i \leq m$ and $1 \leq j \leq n$ are equivalent. Specifically, the optimal solution to (4) occurs when

$$-\frac{\alpha_{i_1}\beta_{j_1}\lambda_{i_1}d_{i_1,j_1}}{(c_{i_1,j_1}-\lambda_{i_1}d_{i_1,j_1})^2} = -\frac{\alpha_{i_2}\beta_{j_2}\lambda_{i_2}d_{i_2,j_2}}{(c_{i_2,j_2}-\lambda_{i_2}d_{i_2,j_2})^2} \quad (7)$$

for $1 \leq i_1, i_2 \leq m$ and $1 \leq j_1, j_2 \leq n$.

It follows that

$$\frac{c_{i,j}-\lambda_i d_{i,j}}{c_{1,1}-\lambda_1 d_{1,1}} = \sqrt{\frac{\alpha_i\beta_j\lambda_i d_{i,j}}{\alpha_1\beta_1\lambda_1 d_{1,1}}} \quad (8)$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Let $\tilde{\lambda}_{i,j} = \alpha_i\beta_j\lambda_i d_{i,j}$. Together with the constraint (5), the set of equations (8) leads to the optimal processing rate allocation as

$$c_{i,j} = \lambda_i d_{i,j} + \frac{\tilde{\lambda}_{i,j}^{1/2}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{\lambda}_{i,j}^{1/2}}(C - \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_i d_{i,j}). \quad (9)$$

Accordingly, the slowdown of a request is

$$s_{i,j} = \frac{\lambda_i d_{i,j}\sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{\lambda}_{i,j}^{1/2}}{\tilde{\lambda}_{i,j}^{1/2}(C - \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_i d_{i,j})}. \quad (10)$$

From (10), we have the following three basic properties regarding the predictability and controllability of service differentiation given by the optimal processing rate allocation scheme:

1) If the session weight or the state weight of a request class increases, slowdown of all other request classes increases, while slowdown of that request class decreases.
2) Slowdown of a request class increases with the increase of session arrival rate and the number of visits to that state of each request class.
3) Increasing the workload (session arrival rate and the number of visits to a state in a session) of a higher request class causes a larger increase in slowdown of a request class than increasing the workload of a lower request class.

Recall $d_{i,j} = v_{i,j}r_j$. From (10), we further have the following service differentiation ratios:

$$\frac{s_{i_2,j}}{s_{i_1,j}} = \sqrt{\frac{\lambda_{i_2}v_{i_2,j}}{\lambda_{i_1}v_{i_1,j}}}\sqrt{\frac{\alpha_{i_1}}{\alpha_{i_2}}} \quad \text{for } j = 1, 2, \cdots, n \quad (11)$$

$$\frac{s_{i,j_2}}{s_{i,j_1}} = \sqrt{\frac{d_{i,j_2}}{d_{i,j_1}}}\sqrt{\frac{\beta_{j_1}}{\beta_{j_2}}} \quad \text{for } i = 1, 2, \cdots, m \quad (12)$$

$$\frac{s_{i_2,j_2}}{s_{i_1,j_1}} = \sqrt{\frac{\lambda_{i_2}d_{i_2,j_2}}{\lambda^{i_1}d_{i_1,j_1}}}\sqrt{\frac{\alpha_{i_1}\beta_{j_1}}{\alpha_{i_2}\beta_{j_2}}}. \quad (13)$$

From (11), (12), and (13), we can see that the optimal processing rate allocation has the property of fairness, as well. That is,

Theorem 3.1: The optimal allocation (9) guarantees relative service differentiation between the requests in both inter-session and intra-session dimensions and their quality spacings with respect to slowdown are square-root proportional to their per-defined differentiation weights.

Remark 1. If session arrival rate λ_i and the resource requirement of a session from a customer class in a state ($d_{i,j}$) are fixed, a request class (i, j) with a higher session weight α_i or with a higher state weight β_j gets more portion of available processing rate of the e-Commerce server. However, we note that the predictability of inter-session service differentiation holds iff $\sqrt{\frac{\lambda_{i_1}v_{i_1,j}}{\lambda_{i_2}v_{i_2,j}}} \leq \sqrt{\frac{\alpha_{i_1}}{\alpha_{i_2}}}$ for all $j = 1, 2, \cdots, n$. Also, the predictability of intra-session service differentiation holds iff $\sqrt{\frac{d_{i,j_1}}{d_{i,j_2}}} \leq \sqrt{\frac{\beta_{j_1}}{\beta_{j_2}}}$ for all $i = 1, 2, \cdots, m$. Otherwise, the essential requirement of 2D service differentiation, predictability, will be violated. The differentiated schedules will be inconsistent. For differentiation predictability, one solution is temporary weight promotion, as suggested in [25]. When it is applied in this context, based on the current session arrival rates and the number of visits to a state in sessions, the scheduler temporarily increases session weights α_i and state weights β_j in the current resource allocation interval so that the predictability of 2D service differentiation holds. In this case, the allocation scheme is heuristic.

Remark 2. We consider the problem of processing rate allocation for 2D service differentiation when constraint (3) holds. Otherwise, a request's slowdown can be infinite and provisioning slowdown differentiation would be infeasible. Session-based admission control mechanisms can be applied to drop sessions from low classes so that constraint (3) holds.

B. A Proportional Slowdown Allocation Scheme

To provide more insights into the impact of various processing rate allocation schemes on 2D service differentiation on e-Commerce servers, we present a proportional share allocation scheme that is tailored from proportional delay differentiation in network packet routing [11], [18].

In general, a proportional resource allocation scheme assigns quality factors to request classes in proportion to their quality differentiation weights. In the 2D service differentiation model for on-line transactions on e-Commerce servers, the quality factors of request classes are represented by their slowdown $s_{i,j}$. Consequently, the proportional allocation model imposes the following constraints for inter-session service differentiation and intra-session service differentiation, respectively:

$$\frac{s_{i_2,j}}{s_{i_1,j}} = \frac{\alpha_{i_1}}{\alpha_{i_2}}, \quad \text{for all } j = 1, 2, \cdots, n \quad (14)$$

$$\frac{s_{i,j_2}}{s_{i,j_1}} = \frac{\beta_{j_1}}{\beta_{j_2}}, \quad \text{for all } i = 1, 2, \cdots, m \quad (15)$$

where α_i and β_i are the normalized quality weighting factors as defined in II-B.

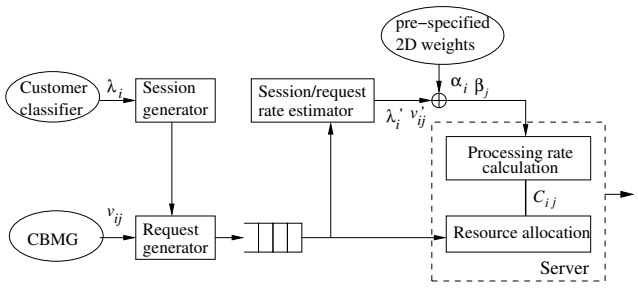


Fig. 1. The architecture of the e-Commerce server simulation model.

Based on the above analyses for the optimization-based processing-rate allocation scheme, we consider the proportional rate allocation problem when constraint (3) holds. If it is violated, 2D service differentiation must be complemented by session-based admission control. Recall the definition of slowdown in (2). According to (14) and (15), we derive the following equation system:

$$\frac{s_{i_2,j_2}}{s_{i_1,j_1}} = \frac{\alpha_{i_1}\beta_{j_1}}{\alpha_{i_2}\beta_{j_2}} \tag{16}$$

for $1 \leq i_1, i_2 \leq m$ and $1 \leq j_1, j_2 \leq n$.

Together with the constraint (5), the set of equations (16) leads to a processing rate allocation as

$$c_{i,j} = \lambda_i d_{i,j} + \frac{\tilde{\lambda}_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{\lambda}_{i,j}}(C - \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_i d_{i,j}). \tag{17}$$

Accordingly, the slowdown of a request is

$$s_{i,j} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{\lambda}_{i,j}}{\alpha_i\beta_j(C - \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_i d_{i,j})}. \tag{18}$$

According to (14) and (15), the proportional allocation scheme generates consistent and predictable schedules for 2D service differentiation on e-Commerce servers. According to (18), it can be verified that this scheme also satisfies the three basic properties of the predictability and controllability achieved by the optimal allocation scheme.

IV. Implementation Issues

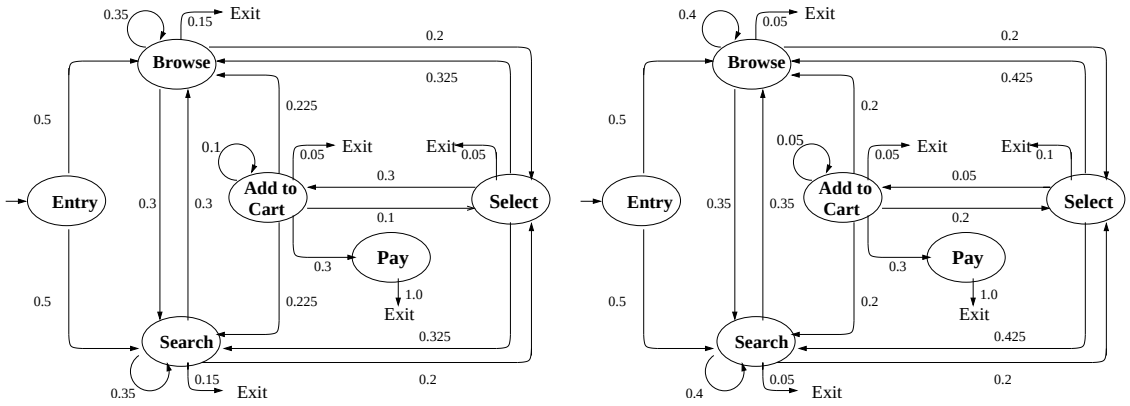
To evaluate the proposed processing rate allocation schemes on provisioning 2D service differentiation, we built a simulation model for e-Commerce servers. We used a synthetic workload generator derived from the real traces [8], [19], [20]. It allowed us to perform sensitivity analyses in a flexible way. Figure 1 outlines the basic architecture of the simulation model. It consists of a customer generator, a session generator, a request generator, a session/request rate estimator, a listen queue, and an e-Commerce server.

Based on the customer classification (*e.g.*, heavy buyer or occasional buyer), the customer generator assigns session arrival rate for each customer class (λ_i). The session generator then produces head requests that initiate sessions for the class. The session generation follows a Poisson process. The subsequent requests of a session are generated by the request generator according to its Customer Behavior Model Graph (CBMG). That is, based on the current state, transition probability and think time associated with each state, requests are generated for the session. Figure 2 shows two CBMGs for a heavy buyer class and an occasional buyer class, respectively. The CBMGs were derived from an on-line shopping store trace [19], [20]. We implemented both profiles in our simulations. Think time between state transitions of the same session is a part of customer behavior. It is generated by an exponential distribution with a given mean.

Each session request is sent to the e-Commerce server and stored in a listen queue. The listen queue is limited to 1,024 entries, which is a typical default value [10]. If the listen queue is full, a new request to the server is rejected and both the request and the whole session is aborted. In the simulations, we simulated a file mix as defined by TPC-W [23], a benchmark of e-Commerce workloads. The average resource demand of sessions in each state is assumed to be the same. It is exponentially distributed with a mean. The e-Commerce server's capacity is 1,000 requests per second for a TPC-W-like file mix and the service time for a request is proportional to the requested file size.

We simulated the proposed processing rate allocation schemes on the e-Commerce server by dividing the request scheduling process into a sequence of short intervals of processing rate calculation and resource allocation. The calculation of processing rate for each class was based on the measured session arrival rate of each class, the number of visits to a state in a session from each class, the average resource demand of a request in a state in a single session, as well as the pre-specified 2D differentiation weights α_i and β_j. A fairly accurate estimation of these parameters is required so that the proposed processing rate allocation schemes can adapt to the dynamically changing workloads. We utilized history information to estimate these values in the session/request rate estimator. The estimate of session arrival rate of each customer class (λ_i') was obtained by counting the number of new sessions from each customer class occurring in a moving window of the past allocation intervals. As a simple way of calculating request arrival rate based on history information, the moving window is widely used in many similar experiments [10]. Smoothing techniques were applied to take weighted averages over past estimates. Similarly, the number of visits to a state in a session from each customer class ($v_{i,j}'$) was estimated.

Like others in [2], [10], [13], [24], we assumed CPU processing rate to be the single resource bottleneck in the e-Commerce server. The interval of CPU processing rate allocation was set to 5 seconds, as CPU utilization is updated on a 5 second interval in some operating systems [10]. The e-Commerce server maintained $m \times n$ listen queues. Given the processing rate for each class $c_{i,j}$ according to the results of (9) and (17), a generalized proportional-share scheduling algorithm (GPS) [15] was simulated to allocate CPU resource between $m \times n$ threads. Each thread processed requests from a request class stored in the corresponding queue.



(a) CBMG for a heavy buyer class. (b) CBMG for an occasional buyer class.

Fig. 2. Customer Behavior Model Graph (CBMG) for different customer classes.

V. Performance Evaluation

In this section, we present some representative simulation results. We considered two customer classes: heavy buyer (class A) and occasional buyer (class B). Their CBMGs are shown in Figure 2 [19], [20]. Because the buy to visit ratio of two customer classes is 0.11 and 0.04, respectively, we assigned session differentiation weight as 11:4 to class A and B. We defined 6 states for each session: pay, add to shopping cart, select, search, browse and entry. They were sorted in a non-increasing order according to their number of state transitions needed to end in the pay state. We assigned state differentiation weight as 5:4:3:2:2:1 to the states correspondingly. The average number of visits to each state in a session is derived from the CBMGs. It is 0.11, 0.37, 1.12, 2.71, 2.71, and 1 for customer class A, and 0.04, 0.14, 2.73, 6.76, 6.76, and 1 for customer class B, respectively. The ratio of session arrival rate of customer class A and B was set to 1:9, according to [19], [20]. Each simulation result is an average of 200 runs.

A. Impact of Service Differentiation on Service Slowdown

Figure 3 shows the results of service slowdown with the increase of server load. Service slowdown is defined to be the optimization objective function in (4). The results were obtained by the use of the optimal and the proportional allocation schemes. For comparison, the figure also includes the results without service differentiation. When the server load is below 10%, slowdown of all request classes is very small. When the server load is above 90%, slowdown of some request classes is very large. Actually, due to the limitation of listen queue size, some requests were rejected and their sessions were aborted. Session-based admission control mechanisms are required when the server is heavily loaded. The focus of this work is on provisioning 2D service differentiation by the use of proposed processing rate allocation schemes. Thus, we varied the server load from 10% to 90%.

Firstly, we found that simulation results agree with the expected results before the server is heavily loaded ($\leq 70\%$). The agreement verifies the assumption made in the modeling (Section II-A) that request arrivals in each state from sessions

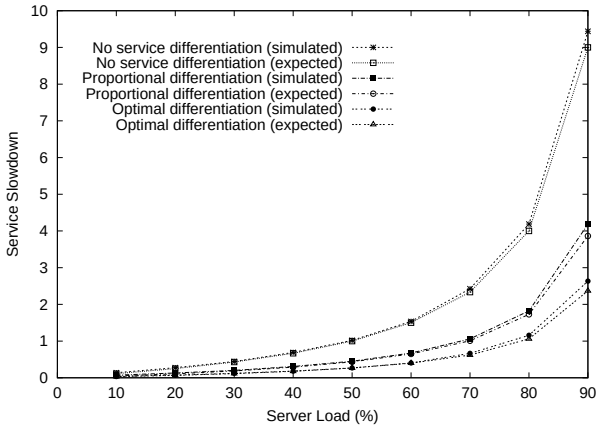
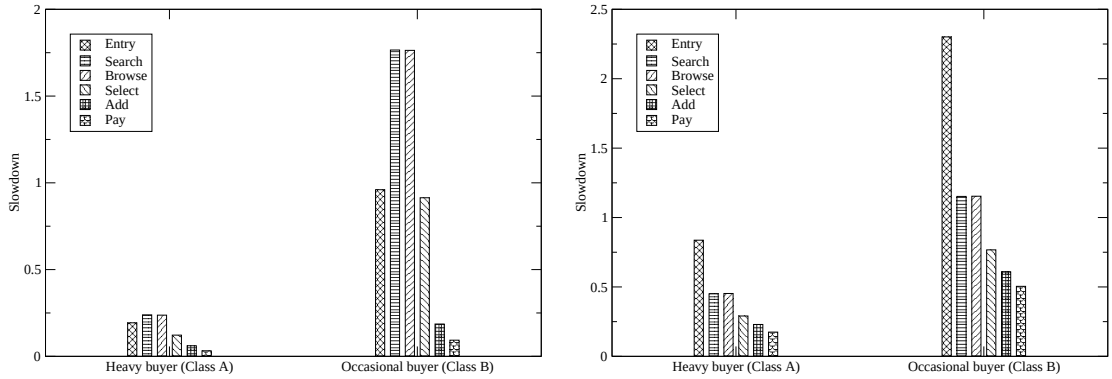


Fig. 3. Service slowdown with the increase of server load.

of a class can be seen to be a Poisson process if session arrivals of that class meet a Poisson process. The gap between the simulated results and the expected ones increases as the server load increases. This is due to the variance of arrival distributions. Secondly, as we expected, the optimal allocation scheme minimizes service slowdown. The proportional allocation scheme achieves higher service slowdown. Obviously, an e-Commerce server without service differentiation provisioning receives much higher service slowdown. In the following, we give more sensitivity analyses.

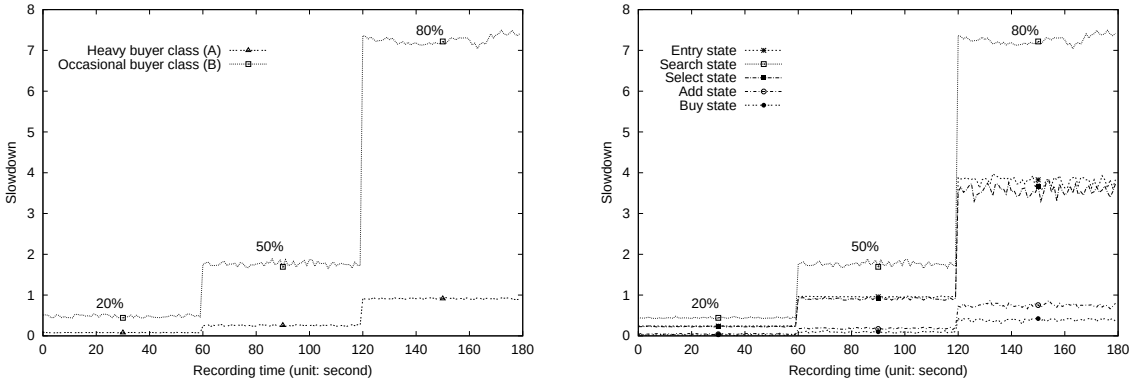
B. Impact of Service Differentiation on Request Slowdown

Figure 4 shows slowdown of individual requests in different states due to inter-session and intra-session service differentiation when the server has 50% (medium) load. The results were due to the optimal allocation scheme and the proportional allocation scheme, respectively. In the figure, each bar shows slowdown of a request class. It can be seen that the requests in the browse state and search state almost have the same slowdown. This is because the customers have similar behaviors at these two states in terms of their state transition probabilities and resource demands. They were assigned the same state weight. In the following discussions, we will use



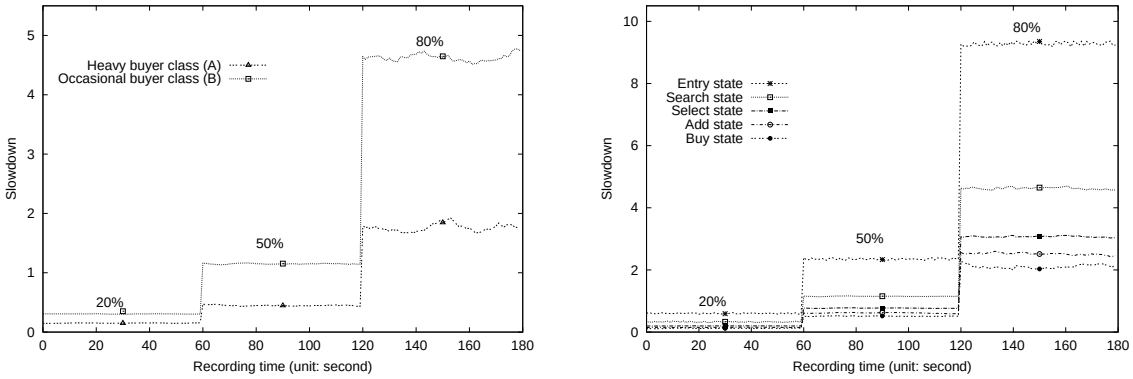
(a) Optimal allocation scheme. (b) Proportional allocation scheme.

Fig. 4. Inter-session and intra-session service differentiation when the server has 50% load.



(a) Inter-session service differentiation (search state). (b) Intra-session service differentiation (class B).

Fig. 5. A microscopic view of slowdown of individual requests due to the optimal allocation scheme.

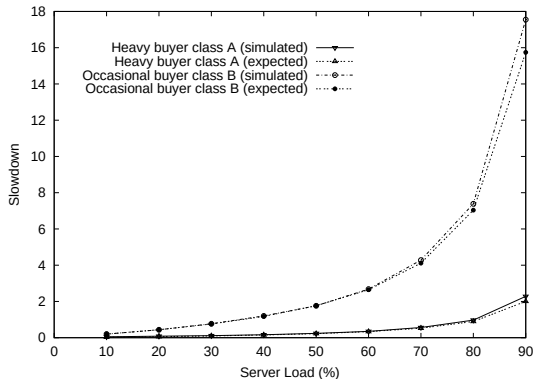


(a) Inter-session service differentiation (search state). (b) Intra-session service differentiation (class B).

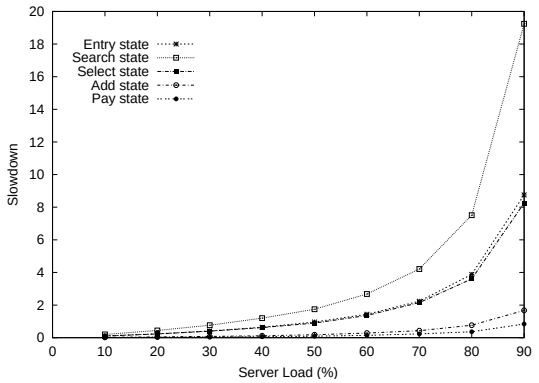
Fig. 6. A microscopic view of slowdown of individual requests due to the proportional allocation scheme.

the search state to represent both and use the results of the search state to address inter-session service differentiation.

Both Figure 4(a) and Figure 4(b) show that the objective of inter-session service differentiation is achieved. In each state, sessions from class A always have lower slowdown than those of class B. From Figure 4(a), it can be seen that the requirement of intra-session differentiation predictability between the entry state and the search state is violated in both class A and B categories. Sessions in the entry state should have higher slowdown than sessions in the search state. Although the optimal allocation scheme minimizes service slowdown, the requirement of $\sqrt{\frac{d_{i,j_1}}{d_{i,j_2}}} \leq \sqrt{\frac{\beta_{j_1}}{\beta_{j_2}}}$ can be violated between the corresponding states, as we discussed in Section III-A. This violation scenario provides an intuition into the fact that the predictability of the optimal processing rate allocations depends on class load distributions. It demonstrates that to provide predictable service differentiation, a scheduler must be able to control the setttings of some parameters (*e.g.*,

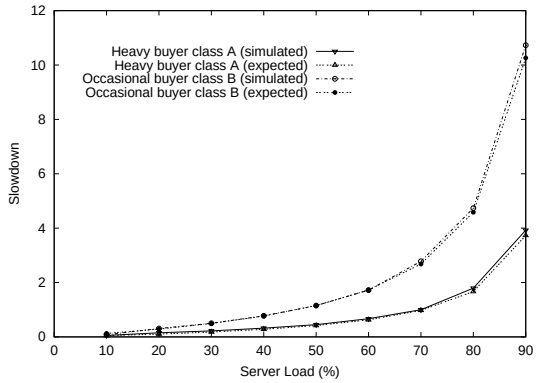


(a) Inter-session service differentiation (search state).

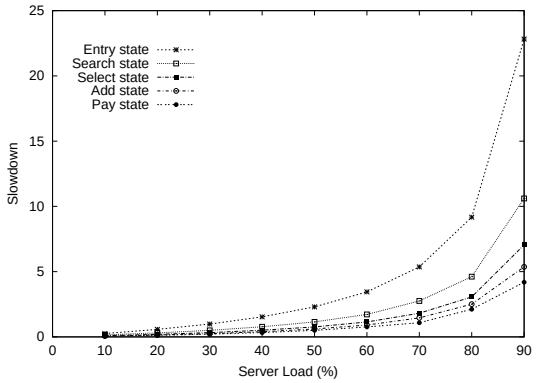


(b) Intra-session service differentiation (class B).

Fig. 7. A long-term view of slowdown of request classes due to the optimal allocation scheme.



(a) Inter-session service differentiation (search state).



(b) Intra-session service differentiation (class B).

Fig. 8. A long-term view of slowdown of request classes due to the proportional allocation scheme.

promoting differentiation weights). By contrast, there are no such violations in Figure 4(b). This is because the proportional allocation scheme guarantees differentiation predictability. The differentiation ratios between request classes are proportional to the ratios of their weights and independent of their load variations. However, this is achieved at the cost of much higher service slowdown, as shown in Figure 3.

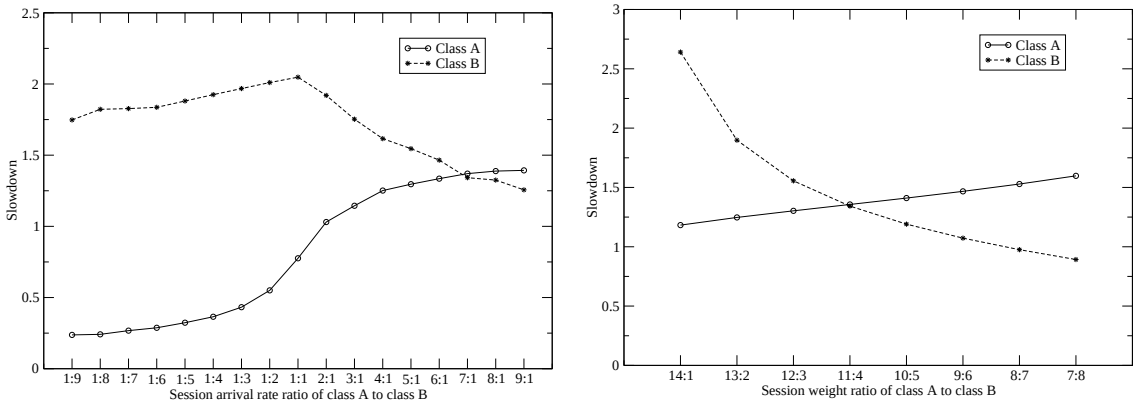
Figure 5 and Figure 6 show a microscopic view of slowdown of individual requests when the server has 20% (light), 50% (medium) and 80% (heavy) load. At each load, the results were recorded for 60 seconds and shown in the figures from left to right. Each point represents slowdown of a request class in consecutive recording time units (seconds). Figures 5(a) and 6(a) illustrate inter-session service differentiation. All sessions are at search state. The plots from other states have the similar shapes. They show that the objective of inter-session differentiation is achieved consistently in the short run. Figures 5(b) and 6(b) illustrate intra-session service differentiation over sessions of class B. The results of class A have similar patterns. In the following, we use results of class B to address intra-session service differentiation.

Figure 7 and Figure 8 show average slowdown of request classes with the increase of server load from 10% to 90%. Figures 7(a) and 8(a) illustrate inter-session service differentiation over sessions in the search state. We have similar results when sessions are in other states. Figures 7(b) and 8(b) illustrate intra-session service differentiation over sessions of class B. From the figures, we can see that both proposed resource allocation schemes can consistently achieve 2D service differentiation at various workloads in the long run.

Figures 7(a) and 8(a) show that the simulated results meet the expectations according to (10) and (18) before the server is heavily loaded ($\leq 70\%$) in inter-session service differentiation. The gap between the simulated results and the expected ones increases as the load increases because the slowdown variance increases. The gap in intra-session service differentiation scenarios has similar shapes. Due to the space limitation, we omit the details in figures 7(b) and 8(b).

C. Impact of Session Arrival Rate and Session Weight

As we explained above, a drawback of the optimal processing rate allocation scheme is the possible violation of differentiation predictability because of the reasons remarked in Section III-A. Figure 9 illustrates violations of inter-session service differentiation at various ratios of session arrival rate and session weight between customer class A and B, when server has 50% load. Due to the space limitation, we omit those of intra-session service differentiation.



(a) Ratio of session arrival rate (class A:B). (b) Ratio of session weight (class A:B).

Fig. 9. Differentiation violations at various ratios of session arrival rate and session weight.

In Figure 9(a), the x-axis shows the ratios between session arrival rates of the two classes (A:B). The session weight ratio is fixed to be 11:4. As the ratio of session arrival rates increases, the gap between the slowdown of class A and that of class B decreases. This is under our expectation according to (11). When the ratio goes up to 7:1, the predictability of inter-session service differentiation becomes violated. To study the effect of session weight promotion, we assumed a fixed ratio of session arrival rate between class A and B as 7:1. Figure 9(a) shows that there is no differentiation violation until the session weight ratio (A:B) goes down to 11:4. The degree of violation increases as the ratio continues to drop.

VI. Related Work

E-Commerce applications are session-based. Current research on scalable e-Commerce servers is mainly on workload characterization [3] and session-based admission control [8], [10]. In [10], Cherkasova and Phaal proposed a group of session-based admission control strategies to prevent a commercial Web server from becoming overloaded. The proposed mechanisms treat sessions from different clients equally and provide a fair guarantee of completion for any accepted session, independent of session length. The performance analyses focused on the throughput gains in terms of completed sessions instead of completed requests when servers were overloaded. In [8], Chen and Mohapatra proposed a dynamic weighted fair scheduling algorithm to control overload in e-Commerce servers. It avoids processing of requests that belong to sessions that are likely to be aborted in the near future. By contrast, the focus of our work is to provision 2D service differentiation when resource demands of workloads are within resource capacity of the server. This is complementary to previous work on session-based workload characterization and admission control for overload protection.

The service differentiation provisioning problem was firstly addressed at the network side. Most previous efforts focused on queueing-delay differentiation in packet level; see [11], [18] for examples. At the server side, a primary focus has been on priority-based request scheduling for responsive time differentiation [2], [5], [9], [12]. For example, in [2], Almeida *et al.* addressed strict priority scheduling strategies for controlling CPU utilization in Web content hosting servers. QoS was introduced by assigning priorities to requests for different contents. Requests of lower priority classes were only executed if no requests existed in any higher priority classes. The results showed that service differentiation can be achieved but the quality spacings among different classes cannot be guaranteed by this kind of strict priority scheduling. Another popular priority scheduling is time-dependent. Time-dependent priority scheduling has been used in achieving queueing-delay differentiation in network communications. It adjusts the priority of a backlogged class according to experienced delays of backlogged packets. Two representative algorithms are WTP [11] and adaptive WTP [18]. This kind of algorithms can be tailored in achieving queueing-delay differentiation at the service side [8], [17]. For example, Chen and Mohapatra recently developed a time-dependent priority-based scheduling strategy, according to the temporal relationship between the requests in a session, to provide different levels of QoS to the requests in different states [8],

Admission control is often used in combination with priority-based scheduling for service differentiation provisioning. In [1], Abdelzaher *et al.* used classical feedback control theory to achieve overload protection, performance guarantees, and service differentiation in Web servers. The strategy was based on real-time scheduling theory which states that response time can be guaranteed if server utilization is maintained below a pre-computed bound. Thus, control-theoretical approaches was formulated to keep server utilization at or below the bound. In [17], Lee et al. proposed admission control algorithms in combination with time-dependent priority scheduling for proportional queueing-delay differentiation on a Web server. Therefore, this kind of admission control by itself is not sufficient for slowdown differentiation provisioning.

In comparison with response time, slowdown is a more accurate performance metric because it is desirable that a request's delay be proportional to its processing requirement. In [13], Harchol-Balter evaluated on-line job assignment

strategies in a distributed server system, where the workload was heavy-tailed and job size was unknown to the scheduler. The primary objective was to minimize mean slowdown of the independent jobs in the distributed system. By contrast, the objective of this paper is to minimize service slowdown, a weighted sum of slowdown of the requests in different sessions and different session states. In [25], Zhu *et al*. used stretch factor, a variant of slowdown, as the performance metric for service differentiation in a cluster of Internet servers. They proposed a demand-driven node partitioning approach for service differentiation provisioning on clusters of servers. The processing rate allocation strategy presented in this paper not only provides 2D service differentiation, but also lends itself to be realizable in various server environments.

The knowledge about customers' navigation patterns is important to provisioning service differentiation on e-Commerce servers. In [19], [20], Menascé *et al*. proposed CBMG to describe customers' navigation patterns through an e-Commerce site. Based on CBMGs, the authors presented a family of priority-based resource management policies for e-Commerce servers [20]. Three priority classes were used: high, medium, and low. Priorities of sessions changed dynamically as a function of state a customer was in and as a function of the amount of money the shopping cart had accumulated. Resource management strategies were geared toward optimizing business-oriented metrics, such as revenue / sec. Their objective was to maximize a global system utility function and the resource allocation strategies cannot control quality spacings among different requests. By contrast, the objective of this paper is to provide predictable, controllable, and fair service differentiation to sessions from different customer classes and to sessions in different states.

VII. Conclusions

There is a growing demand for replacing the current best-effort service paradigm with a model that differentiates requests' QoS based on clients' needs and servers' resource limitations. In this paper, we proposed a 2D service differentiation model with respect to slowdown for session-based e-Commerce applications, namely, inter-session and intra-session service differentiation. We defined a performance metric of service slowdown as weighted sum of slowdown of requests in different sessions and different session states. We formulated the 2D service differentiation model as an optimization problem of the processing rate allocation with the objective of minimizing service slowdown. We derived optimal rate allocations and showed that the optimal allocations guarantee proportional slowdown differentiation between the requests in both inter-session and intra-session dimensions. For comparison, we presented another proportional slowdown allocation scheme that was tailored from proportional queueing-delay differentiation algorithms in networking.

We conducted comprehensive evaluations of the proposed processing resource allocation schemes by the use of proportional-share scheduling. The simulation results have shown that both schemes can consistently achieve 2D service differentiation in the short run and long run. The optimal allocation scheme guarantees 2D service differentiation at a minimum cost of service slowdown.

References

[1] T. F. Abdelzaher, K. G. Shin, and N. Bhatti. Performance guarantees for Web server end-systems: a control-theoretical approach. *IEEE Trans. on Parallel and Distributed Systems*, 13(1):80–96, 2002.

[2] J. Almeida, M. Dabu, A. Manikutty, and P. Cao. Providing differentiated levels of services in Web content hosting. In *Proc. ACM SIGMETRICS Workshop on Internet Server Performance*, pages 91–102, 1998.

[3] M. Arlitt, D. Krishnamurthy, and J. Rolia. Characterizing the scalability of a large Web-based shopping system. *ACM Trans. on Internet Technology*, 1(1):44–69, 2001.

[4] M. A. Bender, S. Chakrabarti, and S. Muthukrishnan. Flow and stretch metrics for scheduling continuous job streams. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[5] N. Bhatti and R. Friedrich. Web server support for tiered services. *IEEE Network*, 13(5):64–71, 1999.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Wang Z., and W. Weiss. An architecture for differentiated services. *IETF RFC 2475*, 1998.

[7] S. Chandra, C. S. Ellis, and A. Vahdat. Application-level differentiated multimedia Web services using quality aware transcoding. *IEEE J. on Selected Areas in Communications*, 18(12):2544–2265, 2000.

[8] H. Chen and P. Mohapatra. Session-based overload control in QoS-ware Web servers. In *Proc. IEEE INFOCOM*, 2002.

[9] X. Chen and P. Mohapatra. Performance evaluation of service differentiating internet servers. *IEEE Trans. on Computers*, 51(11):1,368–1,375, 2002.

[10] L. Cherkasova and P. Phaal. Session-based admission control: A mechanism for peak load management of commercial web sites. *IEEE Trans. on Computers*, 51(6):669–685, 2002.

[11] C. Dovrolis, D. Stiliadis, and P. Ramanathan. Proportional differentiated services: Delay differentiation and packet scheduling. *IEEE/ACM Trans. on Networking*, 10(1):12–26, 2002.

[12] L. Eggert and J. Heidemann. Application-level differentiated services for Web servers. *World Wide Web Journal*, 3(2):133–142, 1999.

[13] M. Harchol-Balter. Task assignment with unknown duration. *Journal of ACM*, 29(2):260–288, 2002.

[14] T. Ibarkai and N. Katoh. *Resource allocation problem - Algorithmic approaches*. The MIT Press, 1988.

[15] J. Kay and P. Lauder. A fair share scheduler. *Communication of ACM*, 31(1):44–55, 1988.

[16] L. Kleinrock. *Queueing Systems, Volume II*. John Wiley and Sons, 1976.

[17] S. C. M. Lee, J. C. S. Lui, and D. K. Y. Yau. Admission control and dynamic adaptation for a proportional-delay DiffServ-enabled Web server. In *Proc. ACM SIGMETRICS*, 2002.

[18] M. K. H. Leung, J. C. S. Lui, and D. K. Y. Yau. Adaptive proportional delay differentiated services: Characterization and performance evaluation. *IEEE/ACM Trans. on Networking*, 9(6):908–817, 2001.

[19] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes. A methodology for workload characterization of E-commerce sites. In *Proc. 1st ACM Conf. on Electronic Commerce*, 1999.

[20] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes. Resource management policies for E-commerce servers. In *Proc. ACM SIGMETRICS Workshop on Internet Server Performance*, 1999.

[21] J. Nielsen. Why people shop on the Web. http://www.useit.com/alertbox/990207.html (Date of access: July 25, 2003).

[22] A. Riska, E. Smirni, and G. Ciardo. ADAPTLOAD: effective balancing in clustered Web servers under transient load conditions. In *Proc. IEEE Int'l Conf. on Distributed Computing Systems (ICDCS)*, 2002.

[23] W. D. Smith. TPC-W: Benchmarking an Ecommerce solution. http://www.tpc.org/tpcw (Date of access: Nov 28, 2002).

[24] T. Zhao and V. Karamcheti. Enforcing resource sharing agreements among distributed server cluster. In *Proc. IEEE Int'l Parallel and Distributed Processing Symposium (IPDPS)*, 2002.

[25] H. Zhu, H. Tang, and T. Yang. Demand-driven service differentiation for cluster-based network servers. In *Proc. IEEE INFOCOM*, pages 679–688, 2001.