Maximizing Service-Level-Agreement Revenues in Clustered-based Web Server Systems

Jian Zhang, Timo Hämäläinen, Jyrki Joutsensalo
Dept. of Mathematical Information Technology
University of Jyväskylä, FIN-40014 Jyväskylä, Finland

***Abstract*— Cluster-based Web server systems have become a major means to hosting e-commerce sites. In this paper, we link the issue of resource partitioning scheme with the pricing strategy in a Service-Level-Agreement (SLA) and analyze the problem of maximizing the revenues attained in the hosting of a e-commerce site with a SLA contract by optimally partitioning the server resources among all supported service classes. The optimal resource partitioning scheme is derived under the linear pricing strategy by the Lagrangian optimization approach, which has the closed-form solution. The simulation results demonstrate the ability of revenue maximization of the derived optimal resource partitioning scheme in cluster-based Web server systems.**

Keywords: Cluster-based Web server systems, QoS, Linear pricing strategy, Optimal resource partitioning scheme, Revenue maximization.

I. Introduction

The Web is changing from a sole communication and browsing infrastructure to an important medium for conducting personal business and e-commerce, which makes the Quality of Service (QoS) an increasingly critical issue. A fundamental characteristic of e-commerce environments is the diverse set of services provided to support the requirements of various businesses and customers, which result in the definition of different service classes. In a typical e-commerce environment, an e-business operator contracts with a Web service provider to provide applications and services to its business customers, which can be consumers (B2C) or other businesses (B2B); in other words, a Web service provider hosts an e-commerce Web site via a contract with the e-commerce operator. In many e-commerce contracts, the Web service provider agrees to offer a certain level of QoS to each class of service in hosting the e-commerce site, and in return the e-business operator agrees to pay the service provider based on the QoS levels received by its customers. These contracts are based on a Service-Level-Agreement (SLA) between the e-business operator and the Web service provider that defines the QoS parameters for each class of service, the anticipated workload intensity of per-class requests from the customers of the e-business and the pricing strategy by which the SLA payment will be determined.

The exponential growth in Internet usage, much of which is fueled by the growth and requirements of various aspects of e-commerce, has created the demand for more and faster Web servers capable of serving over 100 million Internet users. During recent years, server clustering has emerged as a promising technique to build faster, scalable and cost-effective Web servers [9], which makes cluster-based Web server systems become a major means to hosting e-commerce sites. A state-of-the-art cluster-based Web server system consist of a number of back-end server nodes and a specialized front-end node, which acts as the single input point of customer requests and is responsible for distributing the inbound requests among the back-end nodes. The customer requests of an e-business from different service classes share the server resources of a cluster-based Web server system which hosts the e-commerce Web site. In this paper, we analyze the problem of maximizing the revenues attained in the hosting of an e-commerce site with a SLA contract by optimally partitioning the server resources among all supported service classes in the SLA. A number of papers [6], [3], [12], [10], [11] have focused on enabling differentiated services in such cluster-based Web server systems, but none of them addressed the topic of maximizing SLA revenues. The issue of maximizing SLA revenues in the cluster platform of Web server farms was recently studied by Liu *et al* in [7]. A Web server farm is typically deployed to host several Web sites simultaneously on the same platform. Moreover, they assumed that each back-end server node in a Web server farm can serve multiple service classes; then they tried to optimally allocate the resource (e.g., processing capacity) of each server node among its supported service classes to maximize the resulted SLA revenues and the closed-form solution to the optimal resource allocation scheme (i.e., the optimal weights) in each back-end node was not derived in [7]. Whereas, in this paper, we focus on the cluster platform which hosts a single e-commerce site in a cluster-based Web server system. The problem of maximizing SLA revenues in such a Web cluster system is solved by optimally partitioning all the back-end server resources among the supported service classes and the closed-form solution to the optimal resource partitioning scheme for maximizing SLA revenues is derived from the revenue target function by Lagrangian optimization approach.

The rest of the paper is organized as follows. Section 2 first presents our target Web cluster architecture and its queueing model. Then the linear pricing strategy used in this paper is also generally defined there. The closed-form solution to the optimal resource partitioning scheme is derived in Section 3, which can achieve the maximization of SLA revenues in a cluster-based Web server system built upon the target cluster architecture. Section 4 contains simulation part demonstrating the revenue-maximizing ability of the derived optimal resource

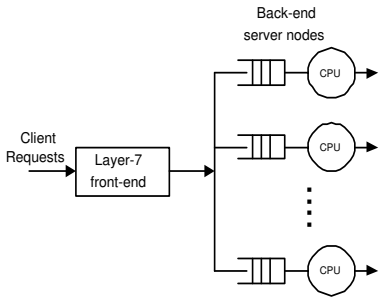

Fig. 1. Queuing model of the target Web cluster architecture.

partitioning scheme. Finally, we present concluding remarks in Session 5.

II. Target Web Cluster Architecture and Linear Pricing Strategy

A. Target Web cluster architecture

Our target Web cluster architecture consists of a front-end component called Web switch and a number of homogeneous back-end server nodes connected by a high-speed LAN. The Web switch acts as the network representative for an e-commerce Web site built upon the target cluster architecture, making the distributed nature of the site architecture completely transparent to the users. In such a way, the authoritative DNS server for the e-commerce site translates the site name into the IP address of its Web switch, which receives all inbound requests destined for the site and then distribute them across the back-end nodes. Moreover, to enable QoS support in the e-commerce site, the Web switch must be able to examine the content of a HTTP request and identify its requested service class, i.e., it is the so-called layer-7 Web switch [9]. The above Web cluster architecture can be further classified on the basis of whether the data from the back-end server nodes to clients (outgoing data) go through the Web switch. In our target cluster architecture, the TCP handoff mechanism [8] is deployed to enable the back-end nodes respond to the clients directly without passing through the front-end nodes as an intermediary. Thus our target cluster architecture can be abstracted as a queuing system shown in Fig. 1.

Several mechanisms [6], [3], [12], [10], [11] have been proposed to enable differentiated services in such a Web cluster system, most of which dynamically partition the server resources among the supported service classes to implement the differentiated QoS levels. In this paper, we link the issue of partitioning server resources among the supported classes with the pricing strategy in a SLA to derive the optimal resource partitioning scheme for maximizing the SLA revenues in hosting an e-commerce site.

Suppose that an e-commerce Web site built upon our target cluster architecture consists of a layer-7 Web switch and *N* homogeneous back-end server nodes, each of which has the processing capacity *C* bits/s; there are totally *m* service classes supported in the site. Then the idea of partitioning server resources among the supported service classes is to partition the *N* back-end server nodes into *m* disjoint server subsets so that each class of requests will be served only by its own assigned server subset. Specifically, the server subset assigned to class *i* is denoted by S_i and the number of server nodes in S_i is denoted by n_i, then $S_i \cap S_j = \oslash$, for $i \neq j$ and $i, j \in [1, m]$, and $\sum_{i=1}^{m} n_i = N$. Thus, the problem of deriving the optimal resource partitioning scheme is actually to find the optimal value of n_i, $i \in [1, m]$. Note that n_i does not have to be an integer, which means that a back-end server node may actually be assigned to multiple service classes with each class taking a portion of it. In this case, we have that back-end node serve those service classes by WFQ algorithm and the WFQ weights equal their shares of that node, respectively.

Based on the analysis in [11], the service time of a class *i* request at a back-end node is proportional to the size of its requested Web object, i.e., $X_i = L_i/C$, where L_i denotes the size (Bytes) of the Web object requested by class *i* customers and *C* (Bytes/s) is the processing capacity of the back-end node. Thus, $\bar{L_i} = E[L_i]$, $\bar{L_i^2} = E[L_i^2]$ and $\bar{X_i} = E[X_i] = \bar{L_i}/C$, $\bar{X_i^2} = E[X_i^2] = \bar{L_i^2}/C^2$. In our scheme, the layer-7 Web switch distributes the inbound requests from class *i* uniformly among the back-end server nodes in server subset S_i to make the server loads balanced. That is to say that, if the overall arrival rate of class *i* requests to the e-commerce site is λ_i requests/s and a back-end server node in S_i is used exclusively by class *i* requests, the mean arrival rate of class *i* requests to that back-end node can be estimated as $lambda_i/n_i$. Furthermore, in this paper, the processing delay at the layer-7 switch is neglected due to the fact that in a Web environment the client-to-server packets are typically much less than the server-to-client packets and the chosen QoS metric in the SLA is the *mean request delay* in the e-commerce Web site. Hence, according to the queuing theory of M/G/1, the analytic mean delay of class *i* requests $\hat{\bar{d}}_i$ in the e-commerce site can be denoted as follows.

$$\begin{aligned} \hat{\bar{d}}_i &= \bar{X}_i + \frac{\frac{\lambda_i}{n_i}\bar{X_i^2}}{2(1-\frac{\lambda_i}{n_i}\bar{X}_i)} \\ &= \frac{\bar{L}_i}{C} + \frac{\lambda_i \bar{L_i^2}}{2C(n_iC-\lambda_i)\bar{L}_i} \end{aligned} \quad (1)$$

The natural constraint of Eq. (1) is $n_iC > \lambda_i\bar{L}_i$ due to the fact that delay can not be negative.

B. Linear Pricing Strategy

As we know, linear and flat linear strategies are among the most used and practical one in real situation. In this paper, our study concentrates on maximizing SLA revenues in the above e-commerce site under the linear pricing strategy and the analysis under the flat pricing strategy is postponed to its sequel. As *mean request delay* is chosen as the QoS metric in the SLA, the linear pricing strategy for class *i* is characterized by the following definition of the linear pricing function $r_i(\bar{d}_i)$.

Definition 1: *The function*

$$r_i(\bar{d}_i) = b_i - k_i\bar{d}_i,\ i = 1, 2, ..., m, b_i > 0, k_i > 0 \quad (2)$$

is called the *linear pricing function* of class *i*, where b_i and k_i are positive constants and $b_i \geq b_j$ and $k_i \geq k_j$ hold to ensure

differentiated pricing if class *i* has a higher priority than class *j* (in this paper, we assume that class 1 is the highest priority and class *m* is the lowest one).

From Eq. (2), it is observed that for any service class, the received SLA revenue by a service provider will decrease linearly along with the increase of offered mean request delay and if the offered mean request delay exceeds the minimum delay requirement of that class, the service provider will obtain a negative revenue, i.e., the service provider will pay the penalties to the e-business operator for failing to meet that minimum requirement. Furthermore, the constant shift b_i determines the maximum price paid for the QoS level received by class *i* requests and the growing rate of penalty depends on the slope k_i.

III. Optimal Resource Partitioning Scheme

Consider an e-commerce Web site built upon the target cluster architecture with *N* back-end server nodes and *m* service classes supported. The processing capacity of each back-end node is *C* bytes/s. The arrival rate of class *i* requests is denoted by λ_i requests/s, $i \in [1, m]$. As the QoS metric considered in the SLA is the *mean request delay*, the mean delay of class *i* requests $\bar{d}_i$ in the e-commerce site will be measured periodically and the SLA revenue due to serving class *i* requests can be determined also periodically based on class *i* pricing function and the above QoS measurement. Specifically, we use Eq. (1) to estimate the real mean delay of class *i* packet $\bar{d}_i$ during one measurement period. Thus, based on the linear pricing function defined in Eq. (2), the SLA revenue *F* obtained in hosting of the e-commerce site during one measurement period is defined as follows.

$$F = \sum_{i=1}^{m} r_i(\bar{d}_i) = \sum_{i=1}^{m}[b_i - k_i(\frac{\bar{L}_i}{C} + \frac{\lambda_i \bar{L_i^2}}{2C(n_iC - \lambda_i\bar{L}_i)})] \quad (3)$$

As a result of the above definition, the issue of maximizing SLA revenue in hosting of an e-commerce Web site can be formulated as follows:

$$max \qquad F = \sum_{i=1}^{m}[b_i - k_i(\frac{\bar{L}_i}{C} + \frac{\lambda_i \bar{L_i^2}}{2C(n_iC - \lambda_i\bar{L}_i)})] \quad (4)$$

$$s.t. \qquad \sum_{i=1}^{m} n_i = N,\ 0 < n_i < N \quad (5)$$

$$n_iC > \lambda_i\bar{L}_i \quad (6)$$

Theorem 1. *Under the linear pricing strategy, the maximum SLA revenue F in hosting an e-commerce site built upon the target cluster architecture is achieved by using the optimal server resource partitioning scheme:*

$$n_i = \frac{(CN - \sum_{j=1}^{m}\lambda_j\bar{L}_j)\sqrt{\frac{k_i\lambda_i\bar{L_i^2}}{2}}}{C\sum_{j=1}^{m}\sqrt{\frac{k_j\lambda_j\bar{L_j^2}}{2}}} + \frac{\lambda_i\bar{L}_i}{C},\ i \in [1, m] \quad (7)$$

Proof: Based on Eqs. (4) and (5), we can construct the following Lagrangian equation.

$$\begin{aligned} P &= P(n_1, n_2, ..., n_m) \\ &= \sum_{i=1}^{m}[b_i - k_i(\frac{\bar{L}_i}{C} + \frac{\lambda_i \bar{L_i^2}}{2C(n_iC - \lambda_i\bar{L}_i)})] + \sigma(N - \sum_{i=1}^{m} n_i) \end{aligned} \quad (8)$$

Set the partial derivative of *P* in Eq. (8) to zero, i.e.,

$$\frac{\partial P}{\partial n_i} = \frac{k_i\lambda_i\bar{L_i^2}}{2(n_iC - \lambda_i\bar{L}_i)^2} - \sigma = 0 \quad (9)$$

It follows that

$$\sigma = \frac{k_i\lambda_i\bar{L_i^2}}{2(n_iC - \lambda_i\bar{L}_i)^2} \quad (10)$$

leading to the solution

$$n_i = \frac{1}{C}(\sqrt{\frac{k_i\lambda_i\bar{L_i^2}}{2\sigma}} + \lambda_i\bar{L}_i),\ i \in [1, m]. \quad (11)$$

Substituting Eq. (11) to Eq. (5), we get

$$\frac{1}{C}\sum_{i=1}^{m}(\sqrt{\frac{k_i\lambda_i\bar{L_i^2}}{2\sigma}} + \lambda_i\bar{L}_i) = N$$

$$\sqrt{\sigma} = \frac{\sum_{i=1}^{m}\sqrt{\frac{k_i\lambda_i\bar{L_i^2}}{2}}}{CN - \sum_{i=1}^{m}\lambda_i\bar{L}_i} \quad (12)$$

And when $\sqrt{\sigma}$ in Eq. (12) is substituted to Eq. (11), the closed-form solution in Eq. (7) is obtained.

Because of the constraint $n_iC > \lambda_i\bar{L}_i$ in Eq. (6), obviously,

$$\sum_{j=1}^{m} n_jC = NC > \sum_{j=1}^{m}\lambda_j\bar{L}_j \quad (13)$$

Hence, the closed-form solution in Eq. (7) $n_i > 0$. Moreover, as $\sum_{i=1}^{m} n_i = N$ and $n_i > 0$, we can conclude that the closed-form solution $0 < n_i < N$.

To prove that the closed-form solution in Eq. (7) is the optimal one, we consider the second order derivative of *P*.

$$\frac{\partial^2 P}{\partial w_i^2} = -\frac{k_i\lambda_i\bar{L_i^2}C}{(n_iC - \lambda_i\bar{L}_i)^3} < 0 \quad (14)$$

due to the constraint in (6). Therefore, the revenue *F* is strictly convex in the interval $0 < n_i < N$, having one and only one maximum. This completes the proof. **Q.E.D.**

Furthermore, when the optimal resource partitioning scheme is deployed, the analytic maximum revenue obtained in hosting the e-commerce site can be acquired by substituting the optimal solution n_i in Eq.(7) into Eq. (3).

IV. Simulations

In this section we present the simulation results which demonstrate the effectiveness of the derived optimal resource partitioning scheme for maximizing the SLA revenues under linear pricing strategy. A number of simulations have been conducted under different parameter settings. In each case, we first numerically determine the optimal resource partitioning scheme using Theorem 1, and then we investigate through simulations the benefits of the optimal scheme by comparing

the SLA revenues obtained under the optimal scheme with those obtained under a natural scheme of proportional resource partitioning as well as the analytic maximum revenues. A representative set of these simulations are presented herein. Throughout this section, we shall focus on an e-commerce Web site consisting of a layer-7 Web switch and 16 back-end server nodes (N=16), where the processing capacity of each back-end node C equals 5.95MB/s and the number of service classes supported in the site $m = 3$ (namely, Gold, Silver and Bronze classes).

For actual Web workloads, it is recognized that Web object sizes are distributed with a heavy tail. Here the Bounded Pareto distribution ($BP(p, q, \alpha)$) [5] is used to model the heavy-tailed characteristic of Web objects. Specifically, the mean size of Web objects is set to 21KB as measured in [2] and p=1KB and q=10MB are chosen as the reasonable minimum and maximum Web object size, respectively. The resulting α=0.8037 is within the range of α values measured in [1] and [4]. The arrival process of client requests destined for the e-commerce site was modelled by Poisson distribution. Additionally, we first set the base arrival rates for each service class and then a multiplicative *load factor* $\rho > 0$ is used to scale these base arrival rates to consider different workload intensities; i.e., $\lambda_j \rho$ will be used in the simulations as class-j arrival rate. The base arrival rate for Gold, Silver and Bronze classes is 100 requests/s, 150 requests/s and 250 requests/s, respectively, throughout the following simulations. As mentioned above, we deploy a scheme that partitions the server resources among the supported service classes in proportional to their inbound workload ($\lambda_i \bar{L}_i$) (bytes) for comparison with our derived optimal resource partitioning scheme, which was also used by Liu *et al* in [7]. Specifically, the *proportional* scheme results in the number of server nodes assigned to class i as below: $n_i = \frac{\lambda_i \bar{L}_i}{\sum_{j=1}^{m} (\lambda_j \bar{L}_j)}, i \in [1, m]$. Note that this proportional scheme is a natural way to allocate server resources in a Web cluster system.

A. The first set of simulations

In the first set of simulations, the set of linear pricing functions deployed is as follows: for the Gold class, $r_1(\bar{d}_1) = 200 - 5000\bar{d}_1$, for the Silver class, $r_2(\bar{d}_2) = 120 - 2000\bar{d}_2$ and for the Bronze class, $r_3(\bar{d}_3) = 40 - 500\bar{d}_3$ (note that the time unit is second here). We first investigate the relationship between the mean request delay generated by simulation and the analytic mean request delay by Eq. (1) (the closed-form solution in Eq. (7) is substituted into Eq. (1)) under different workload intensities. Then the simulation-generated SLA revenue by the optimal resource partitioning scheme is evaluated by comparing it with the one by the proportional scheme as well as the analytic maximum SLA revenue under multiple workload levels.

Fig. 2 shows the simulation-generated mean request delays by the optimal scheme and the analytic mean request delays under different load factors. Fig. 3 presents the simulation-generated request delays by the proportional resource partitioning scheme under the same workload intensities. It can be seen from Fig. 2 that for each service class, its simulation-

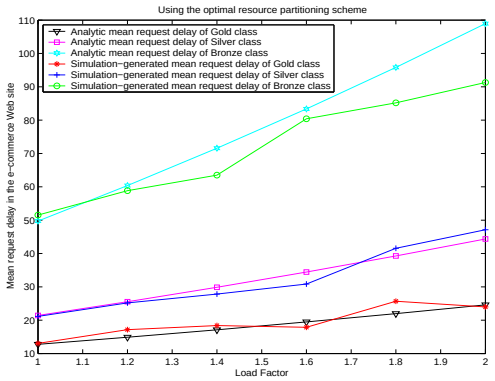

Fig. 2. For the first set of simulations: the mean request delays by the optimal scheme under different load factor ρ.

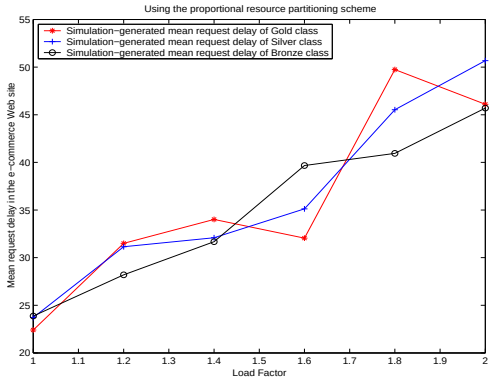

Fig. 3. For the first set of simulations: the mean request delays by the proportional scheme under different load factor ρ.

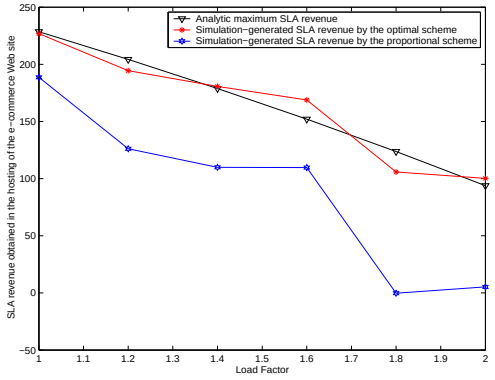

Fig. 4. For the first set of simulations: SLA revenue comparison under different load factor ρ.

generated mean request delay is always pretty close to its analytic mean request delay under all the offered load factors, which demonstrates the correctness of the above assumption that the real mean request delay of class i, $i \in [1, m]$, can be estimated by Eq. (1). Additionally, the optimal resource partitioning scheme enables the differentiated services in the e-commerce Web site as shown in Fig. 2, whereas the proportional scheme does not.

Fig. 4 shows the simulation-generated SLA revenues by the optimal scheme, the simulation-generated revenues by

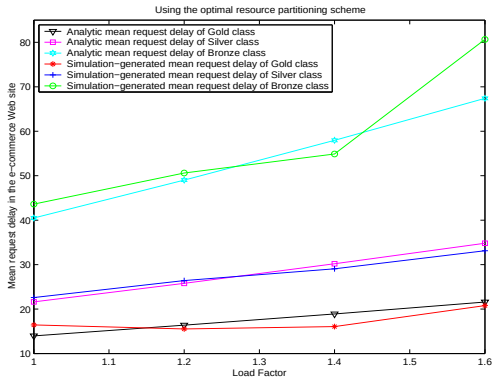

Fig. 5. For the second set of simulations: the mean request delays by the optimal scheme under different load factor ρ.

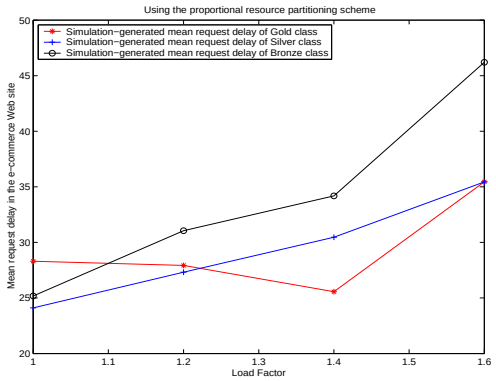

Fig. 6. For the second set of simulations: the mean request delays by the proportional scheme under different load factor ρ.

the proportional scheme and the analytic maximum revenues, respectively, under the same load factors. It is clear that the optimal resource partitioning scheme achieves the maximum revenue under different workload intensities. More importantly, the value of the simulation-generated revenue by the optimal scheme is very close to the one of the analytic maximum revenue, which demonstrates the effectiveness of the derived optimal resource partitioning scheme for maximizing SLA revenues.

B. The second set of simulations

In this part, the above simulations were made again under different linear pricing functions to investigate the performance robustness of the optimal resource partitioning scheme for maximizing SLA revenues. The set of linear pricing functions deployed in the second set of simulations is as below: for the Gold class, $r_1(\bar{d}_1) = 200 - 10000\bar{d}_1$, for the Silver class, $r_2(\bar{d}_2) = 150 - 5000\bar{d}_2$ and for the Bronze class, $r_3(\bar{d}_3) = 80 - 2000\bar{d}_3$. Figs. 5-7 present the simulation results, where it is clear that the simulation-generated mean request delay by the optimal scheme is very close to the analytic mean request delay and the simulation-generated SLA revenue by the optimal scheme, which is larger than the one by the proportional scheme, is always pretty close to the analytic maximum SLA revenue although different set of linear pricing functions is deployed. Therefore, we conclude that our derived optimal resource partitioning scheme can succeed in implementing the maximization of SLA revenues under different workload intensities and different set of linear pricing functions when a Web service provider hosts an e-commerce Web site by cluster-based Web server systems.

V. Conclusions

In this paper, we link the issue of resource partitioning scheme with the pricing strategy in a Service-Level-Agreement (SLA) and explore the problem of maximizing the SLA revenues in the hosting of an e-commerce Web site with a SLA contract by optimally partitioning server resources among the supported service classes. The optimal resource partitioning scheme is derived under the linear pricing strategy, which has the closed-form solution to the optimal number

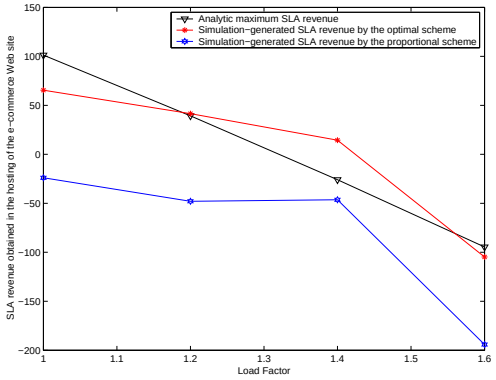

Fig. 7. For the second set of simulations: SLA revenue comparison under different load factor ρ.

of the back-end server nodes assigned to each service class. The simulation results demonstrate that the derived optimal resource partitioning scheme can succeed in implementing the maximization of SLA revenues under different system parameter settings when a Web service provider hosts an e-commerce Web site by cluster-based Web server systems.

In the future work, the issue of revenue maximization under flat pricing strategy will be investigated.

References

[1] M. F. Arlitt and T. Jin, "A workload characterization study of the 1998 World Cup Web site," IEEE Network, 14(3):30-37, May/June 2000.

[2] M. Arlitt and C. Williamson, "Web server workload characterization: the search for invariants," In Proceedings of ACM SIGMETRICS, pp. 126-137, 1996.

[3] V. Cardellini, E. Casalicchio, M. Colajanni, and S. Tucci, "Mechanisms for quality of service in web clusters," Computer Networks, Elsevier Science, Vol. 36, No. 6, pp. 759-769, Nov. 2001.

[4] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," IEEE/ACM Trans. on Networking, 5(6):835-846, Dec. 1997.

[5] M. Crovella, M. Harchol-Balter, and C. Murta, "Task assignment in a distributed system: improving performance by unbalancing load," Performance Evaluation Review, Vol. 26, No. 1, pp. 268-269, 1998.

[6] V. Kanodia and E. W. Knightly, "Multi-class latency-bounded web services," In Proc. Int'l Workshop on Quality of Service, June 2000.

[7] Z. Liu, M. Squillante, and J. Wolf, "On Maximizing Service-Level-Agreement Profits," In Proceedings of the 3rd ACM conference on Electronic Commerce, pp. 213C223, 2001.

[8] V. S. Pai, M. Aron, G. Banga, M. Svendsen, P. Druschel, W. Zwaenepoel, and E. Nahum, ”Locality-Aware Request Distribution in Cluster-based Network Servers,” In Proceedings of the 8th Conference on Architectural Support for Programming Languages and Operating Systems, Oct. 1998.

[9] T. Schroeder, S. Goddard, and B. Ramamurthy, ”Scalable Web server clustering technologies,” IEEE Network, 14(3):38-45, May/June 2000.

[10] J. Zhang, T. Hämäläinen and J. Joutsensalo, ”Dynamic Partitioning: A Mechanism for Supporting Differentiated Services in Cluster-based Network Servers,” In Proc. of 16th Nordic Teletraffic Seminar (NTS16), pp. 185-193, August 21-23, 2002.

[11] J. Zhang, T. Hämäläinen and J. Joutsensalo, ”A New Mechanism for Supporting Differentiated Services in Cluster-based Network Servers,” In Proc. of 10th IEEE/ACM International Workshop on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS2002), pp. 427-432, Oct. 12-16, 2002.

[12] H. Zhu, H. Tang, and T. Yang, ”Demand-driven service diffentiation in cluster-based network servers,” In Proc. of IEEE Infocom2001, Apr. 2001.