International Journal of
Electronic Commerce Studies
Vol.2, No.1 , pp.19-34, 2011

COST ESTIMATION QUEUING MODEL FOR LARGE-SCALE FILE DELIVERY SERVICE

Chih-Chin Liang
National Formosa University
lgcwow@gmail.com

Hsing Luh
National Cheng-Chi University
slu@nccu.edu.tw

ABSTRACT

A large-scale content delivery service within a company in Taiwan uses 40 servers to deliver files from the company's ordering system. In a content delivery system, servers and connections are the resources for sending files. However, because of increasing business needs, the content delivery service must be used for transferring various files rather than just a single business unit. This increases the content delivery frequency. Once a connection is used for delivering a file, the occupied connection serves no other file delivery purpose. Because each connection has a cost, the use of a connection that delivers no other packets represents an opportunity loss, with the necessity of delivering a higher number of packets creating a higher opportunity loss for the existing service. Further, each file sent through this content delivery service has its specified importance because of the different business needs. The average loss caused by the failure to send an important file is considerably larger than that when sending a regular file. That is, the opportunity costs of sending various types of files are different. Because the connections are sparse resources, it is important to discuss a way to use these connections efficiently for delivering various files. In other words, the manager must find a way to control the loss caused by the content delivery service because the opportunity cost of failing to deliver content with different priorities must be endurable. Therefore, this work analyzes the cost function of the queuing model of the proposed content delivery service. The results show that a manager can control the reservation of connections in order to deliver important packets with a minimum cost. That is, the manager can manage content delivery as efficiently as possible.

Keywords: cost estimation, content delivery

1. INTRODUCTION

The large-scale service sector regards uninterrupted customer service as one of the most important operational goals[1]. A typical large-scale service company has various devices such as personal computers that are used by clerks for serving customers. Each device hosts different client software applications that are linked to the systems in the back offices. Obviously, irrespective of the business architecture that is adopted, data will be exchanged among various devices. For example, the case company has an ordering system and a billing system, and there is a need to exchange data between them in order to complete service operations[2, 3]. Previous studies have described distribution failures, where a single server was used to update the contents of all of the clerks' devices; it has since been learned that disasters can halt a content delivery service. Previous studies reported a reliable approach for delivering content along an organizational hierarchy called a Fire and Forget Distribution System (FnFDS), which distributes the updated content to more than 20,000 destinations robustly[4]. Previous studies analyzed FnFDS on the basis of the number of servers and connections to estimate the reliability and efficiency of the system[5, 6]. The results showed that FnFDS requires only 19 servers to meet the content delivery requirement and exhibit outstanding performance; the failure rate of the content delivery service was 1.90735 occurrences per million hours, which is better than the six-sigma capability: 3.4 occurrences per million hours[7, 8, 9]. Obviously, the use of 19 servers in a company environment is an optimal solution for deploying applications.

However, a company needs to consider not only reliability but also the business requirements because of the focus on the cost/benefit ratio. That is, for content delivery, a server needs to serve not only the ordering system but also other business systems/units. Obviously, to run every type of business system/unit, all of the connections from a server to the destinations must be used heavily. Once a connection is selected for delivering a file, no other packets can be transmitted through it. However, the connection must be used as efficiently as possible. Moreover, each file has its importance; if an important file is not delivered to its destination because all of the connections are occupied by other files, the loss from this delivery failure will be significant. Therefore, if the FnFDS is used to deliver every type of content, the manager must measure the loss and efficiency cost. Further, in order to serve more customers, along with adopting cloud computing for the example company, the system scale of the content delivery must be enlarged [10, 11, 12, 13, 14, 15]. That is, 19 servers are insufficient to satisfy the need to serve more customers in the cloud. In such a complicated and large content delivery system, it is important to discuss a way to find an optimal resource

allocation method. If a file needs to be delivered to a large number of destinations through a large-scale network, the resource allocation cost function must be derived. If a manager does not want to bear any additional cost for deploying an application, the application should be withdrawn. By using a measurable cost function, the manager can decide on how to efficiently use the resources owned by the company[5]. However, very few previous studies have discussed the cost function of direct content delivery within an organization[1].

Previous networking studies have shown that through Differentiated Service (DiffServ), a manager can provide different services by using networking resources. With respect to efficient content delivery, previous studies on DiffServ have proposed solutions based on a scheduling program [16, 17, 18]; by assigning a privilege to the packets on the router, each package to be transmitted to its destination(s) will be queued up in an ordered manner according to its privilege. However, this solution is not feasible for business because a file can be sent from anywhere. A router can assign a priority to each file to be sent, but this priority reveals nothing about how important the file is to the business. For example: Is a regular file issued by the Chief Executive Officer (CEO) more important than an urgent file sent by a project manager? That is, is a file important based on its sender? A file is important when it is related to the core business; the other files have low priority. For example, the promotion code of the ordering system is important, but the daily order assigned by the manager has low priority. If the promotion code is not sent to all of the client devices within a specified time interval, customer service might be halted because of inconsistent information. Once the daily order that the manager has issued is sent, even with a delay, the business is still workable. Further, the loss of failing to deliver important packets is considerable, but the frequency of sending important files is low. Regular files also need to be delivered through the connections, but they can wait until after the important files are sent. The assignment of privilege on the router is simply a matter of assigning the order of each packet, irrespective of its business importance. Therefore, it is insufficient to merely assign priorities to packets for scheduling their transmission through the router. Another method to control the use of connections is the use of different algorithms for delivering files, such as the pull-based and push-based content delivery methods. If the packets are sent by using a pull-based method, all of the clients retrieve files from the servers; the server connections are controlled by the clients. If the files are sent using push-based applications, the server will actively deliver packets to the clients; the server connections are controlled by the server. However, the abovementioned two content delivery methods merely discuss the control of the server connections. They are not related to the business, but

their fault tolerance and efficiency are a major concern[5, 6]. However, the manager of the example company needs to know the cost of the connections in order to determine the optimal use of resources. Yet, the abovementioned solutions are infeasible for devising a cost function. It is difficult to measure the cost of each connection through DiffServ or the abovementioned content delivery methods because the cost is computed indirectly. By using an indirect method, we can find the cost of a connection from the importance of each file, but the files are issued from multiple locations and the importance of each file is difficult to quantify. Therefore, the manager finds it difficult to control the connection cost.

In order to directly manage the connections, the reservation of a connection could be adopted. Previous studies have shown that a manager can reserve the use of partial connections for delivering important packets instead of controlling the privileges of packets. Through reservation, a manager can derive a connection usage cost function and manage important and regular packets[19, 20, 21]. In order to derive the cost function, in this study, we built a queuing model to determine the relationship between file transmission and connections. By using the proposed cost function, the manager can reserve connections to deliver important packets and must decide how many connections are required.

The rest of this paper is structured as follows. Section 2 describes the use of FnFDS connections in detail. Section 3 describes the queuing model for the links of the assumed 125 servers. Next, Section 4 discusses the cost of the queuing model. Finally, a discussion and concluding remarks are presented in Section 5.

2. SYSTEM DESCRIPTION

In order to derive the cost function, the system architecture must be introduced. On the basis of a real networking environment, the topology of an FnFDS network is assumed to be a completed network; each node has full links to other nodes within the same networking segment[10, 11]. Further, although FnFDS is online, the servers and links are used, not just for the content delivery service of the ordering system but also for every type of packet delivery with the evolution of the business, including the delivery of mail, emergency information, daily updates, and instant messages. In other words, the network is busy. Furthermore, the packets can be delivered not only to all nodes but also to partial nodes, depending upon the requirement. For example, if a message needs to be delivered to the northern business unit, the message will be delivered to no other business units. Moreover, each undelivered packet has to be re-sent to the destinations immediately

once a connection error occurs. However, blocked packets cannot be re-sent multiple times. That is, the capacity of the sender is limited with respect to the wait time for the packets to be sent. Without loss of generality, in this study, we assume that the queue for re-sending the packets is limited. By using FnFDS, a manager mainly delivers packets through the root server, but he/she can also use the other servers to send files. That is, each server can assume the role of a root server. Therefore, once a server receives a sent file, the server will deliver the file to other linked servers.

From the above description, it can be concluded that in terms of cost efficiency, the manager must understand how to use the connections of a server (sender) efficiently. Important packets (urgent files) must be delivered on time; however, the network might be occupied with the transmission of other packets (regular files). Therefore, partial connections must be reserved for sending urgent files. Once there are K outbound lines of the root server, the cost analysis of this study attempts to determine how many lines of K are required for transmitting urgent files. That is, this work attempts to set up a signal for the remaining lines to deliver urgent files. For example, if the signal is assigned to "N" connections, the connections will be reserved for delivering urgent packets only when "N" and above connections are occupied.

Further, if the important packets cannot be delivered, the business cannot function appropriately. For example, if a "system shutdown" notice cannot be sent to all of the receivers that need this information, the business operation will have to be terminated. Moreover, the frequencies of sending packets during peak hours and regular hours are different. The term "peak hour" refers to the period during which the managers send messages frequently, and therefore, packets will be sent more often during the peak hours than during the regular hours.

The actual business operations are described as follows: On average, each file (packet) should be delivered to all destinations within 5.5 min. The content delivery frequency is set to 5 times the general transmission rate per minute and 2 times the important/emergency packet delivery rate per minute during the peak hours. The content delivery frequency is set to 0.06 times the general transmission rate per minute and 0.01 times the important/emergency packet delivery rate per minute during the regular hours. The regular packets that were blocked because all of the connections were occupied at a certain time should be re-sent. That is, these blocked packets enter a retrial group to wait for available network connections. A blocked packet in the retrial group will be re-sent after the network is available for transferring packets. Moreover, the retrial group is finite because of the sparse business resources. Therefore, it is possible that the

message would be discarded. In the case of the example company, the probability of failing to re-send packets is almost 30% during the peak hours and 5% during the regular hours.

Finally, because the cost of failing to deliver packets is difficult to illustrate, we make an assumption in this study. In order to make this assumption, we interviewed some managers who are in charge of content delivery. From these interviews, we found that the cost of blocking important packets is 20 times the cost of holding packets; the cost of blocking regular packets is 2 times the cost of blocking regular packets.

3. MODEL

This work investigates a queuing model for deriving the cost function of FnFDS. There are two types of content: regular files and urgent files. Let λ_r and λ_u be the average arrival rates of regular files and urgent files, respectively. On the basis of the described scenario, this work assumes that there are K links (connections or transmission channels) in FnFDS, where $1 < K < \infty$. The link transmission rate (service rate) of both types of files (regular files and urgent files) is μ. If there exists at least one free link upon the arrival of urgent files, it is immediately available for data transmission. Otherwise, the attempt to transmit urgent files is forcefully terminated. Let P_u be the probability that urgent files cannot be transmitted (do not receive service).

When there are N $(1 \leq N \leq K)$ or more links that are busy, any incoming regular files are blocked and enter the retrial group. The system attempts to re-send the blocked regular files from the retrial group after the retrial time, which is exponentially distributed at rate α. This work assumes that the capacity of the retrial group, R, is finite R. This assumption is not unusual with respect to intra-network content delivery services. Let P_r be the probability that regular files cannot be transmitted (do not receive service) on the first attempt. Moreover, let H_0 be the probability that blocked regular files enter the retrial group. Then, $1 - H_0$ is the probability that regular files are blocked and leave the system forever. The retrial rate of the blocked files in the retrial group is α. Let H_1 be the probability that regular files in the retrial group return to the retrial group after an unsuccessful retrial. Then, $1 - H_1$ is the probability that the regular files in the retrial group leave the system after an unsuccessful retrial. The average number of regular files in the retrial group is L.

In this study, we investigate the total expected cost function of this file transmission system. Our objective is to minimize the total expected

opportunity cost by setting the threshold N when the number of links is K. In this study, we devise a total expected opportunity cost function (per unit time) for the *N*-signal $M/M/K$ queuing system, where N is the decision variable. The total cost function (per unit time) can be expressed as

$$TC(N,K) = C_h L + C_r P_r + C_u P_u, \qquad (1)$$

where C_h is the holding cost per unit time for each file present in the system, C_r is the blocking cost for each lost regular file, and C_u is the blocking cost for each lost urgent file.

Let $X_1(t)$ be a random variable that represents the number of blocked regular files in the retrial group at time t. Let $X_2(t)$ be a random variable that represents the number of files under transmission (in service) at time t, including regular files and urgent files. Then, $X_1(t)$ and $X_2(t)$ can be modeled as a 2-dimensional Markov chain, whose Q-matrix can be given by

$$q((i,j),(i',j')) = \begin{cases} \lambda_r + \lambda_u & \text{if } i' = i,\ j' = j+1, 0 \le i \le R, 0 \le j \le N-1 \\ \lambda_u & \text{if } i' = i,\ j' = j+1, 0 \le i \le R, N \le j \le K \\ i\alpha & \text{if } i' = i-1,\ j' = j+1, 1 \le i \le R, 0 \le j \le N-1 \\ i\alpha(1-H_1) & \text{if } i' = i-1,\ j' = j, 1 \le i \le R, N \le j \le K \\ \lambda_r H_0 & \text{if } i' = i+1,\ j' = j, 0 \le i \le R, N \le j \le K \\ j\mu & \text{if } i' = i,\ j' = j-1, 0 \le i \le R, 1 \le j \le K \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Let $P(i,j)$ be the stationary probability at state (i,j). Then, the balance equation can be derived as follows:

$$\sum_{i=0}^{R}\sum_{j=0}^{K} P(i,j)\, q\left((i,j),(i',j')\right) = 0 \qquad (3)$$

By solving the balance equation, this work obtains the stationary probability $P(i,j)$. Therefore, some important performance measures of this queuing system can be determined as follows: The average number of regular files in the retrial group is

$$L = \sum_{i=1}^{R}\sum_{j=0}^{K} iP(i,j) \qquad (4)$$

The probability that regular files cannot be transmitted on the first attempt is

$$P_r = \sum_{i=0}^{R}\sum_{j=N}^{K} P(i,j) \qquad (5)$$

The probability that urgent files cannot be transmitted (do not receive service) on the first attempt is

$$P_u = \sum_{i=0}^{R} P(i, K) \qquad (6)$$

4. DISCUSSION

In order to understand the cost function of the proposed queuing model, this work uses the parameters from an actual case. From the descriptions and data given in previous studies[9, 10, 11], this study considers the following parameters:

μ = 0.5, 1, and 1.5 (service rate, 1/μ implies that each packet set should spend 1/μ min to deliver content to all destinations. The number "1" is the current value under 19 servers).

$R = 20$ (capacity of the retrial group),

K = 124 (transmission channels; in this case, we assumed that 125 servers carry out the content delivery),

$H_0 = 1$ (probability that blocked regular files enter the retrial group),

$H_1 = 0.95$ (probability that regular files in the retrial group return to the retrial group after an unsuccessful retrial),

$C_h = 1$ (holding cost per file),

$C_r = 2$ (blocking cost for regular files),

$C_u = 20$ (blocking cost for urgent files),

$\lambda_r = 93$ (files/min)/$\lambda_u = 31$ (files/min) during peak hours, and

$\lambda_r = 46.5$ (files/min)/$\lambda_u = 15.5$ (files/min) during regular hours.

However, to validate the proposed model, this work applies the above derived model to the FnFDS. Based on the real situation, we use a service rate of 1, and the system attempts to re-send a file to its destination every 6 s. The results are shown in Fig. 1. The managers found that, based on the experimental results, the optimal reservation N = 9 and the MC will increase from 30.75 to 30.76 when N = 10. The above result is meaningful to managers, because the costs C_h, C_r, and C_u were assigned based on a survey of managers. These results were verified by managers in a second survey after the model and results were derived.

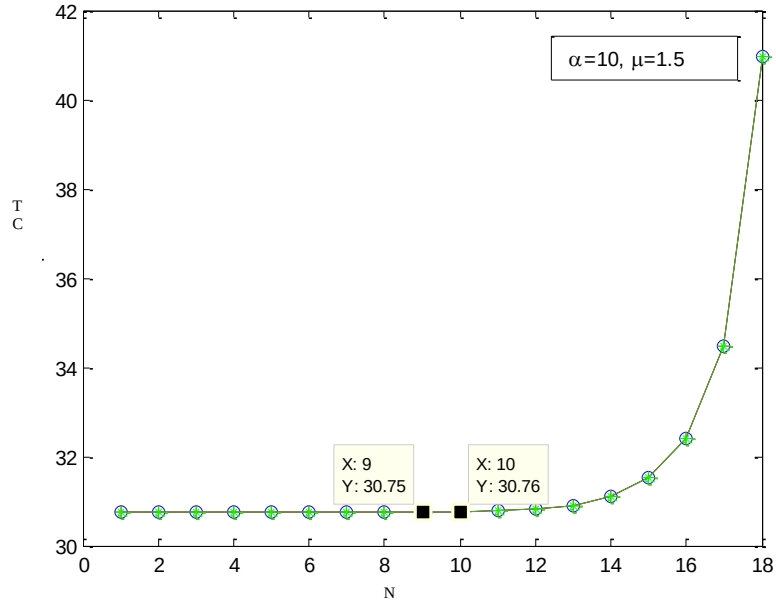

Figure. 1 Threshold N versus managed cost $TC(N, 18)$ and μ = 1 during peak hours.

Further, in order to compare the differences in the frequency of retrying to send a file to its destination, this work assumes the following three parameters:

$\alpha = 10, 60$ and 120 (α = 10 implies that the system attempts to re-send a file to its destination every 6 s).

Figs. 2-7 show diagrams of the managed cost function (MC) along with *N*. Fig. 1 illustrates MC with a service rate of 0.5 during peak hours. A large value of α implies a large value of MC. This in turn implies that the cost along with the decreased reservation for important files (*N* is increasing) will be increased rapidly once the important files are heavily blocked.

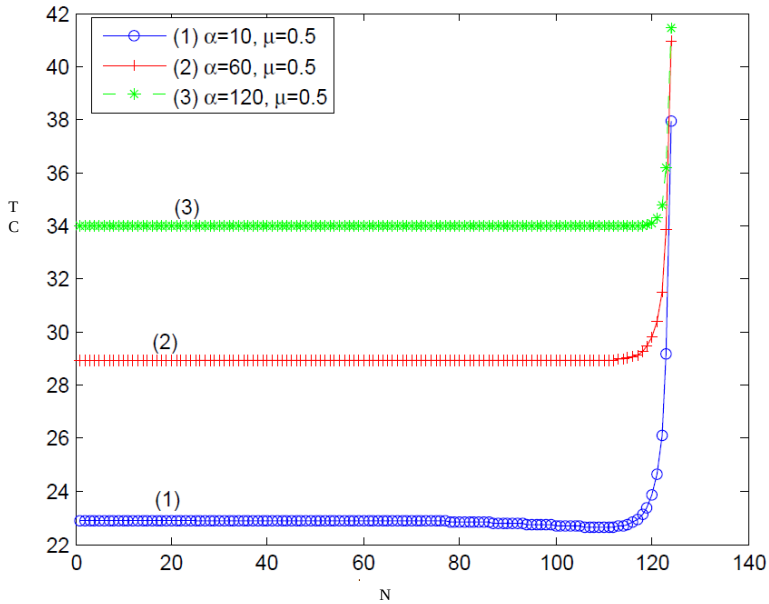

Figure. 2 Threshold N versus managed cost $TC(N, 124)$ and μ = 0.5 during peak hours.

Fig. 2 illustrates MC with a service rate of 1 (faster than the service rate of 0.5) during peak hours. The cost along with the reservation for important files (N is increasing) is decreased at first and will be increased rapidly once the important files are heavily blocked. This figure also shows that the minimum cost can be calculated. The minimum costs are (N = 122, α = 10), (N = 117, α = 60), and (N = 112, α = 120). Moreover, along with different retrial rates (α), the costs change with different shapes: in area I, the order of cost (MC_α) is $MC_{120} > MC_{10} > MC_{60}$, but in area II, this order is $MC_{120} > MC_{60} > MC_{10}$. This is because of the cost function ($MC(N, K) = C_h L + C_r P_r + C_u P_u$). In area I, the increasing C_h is larger than the increasing C_u; therefore, $MC_{10} > MC_{60}$. In area II, because the increasing C_u is larger than the increasing C_h, $MC_{60} > MC_{10}$.

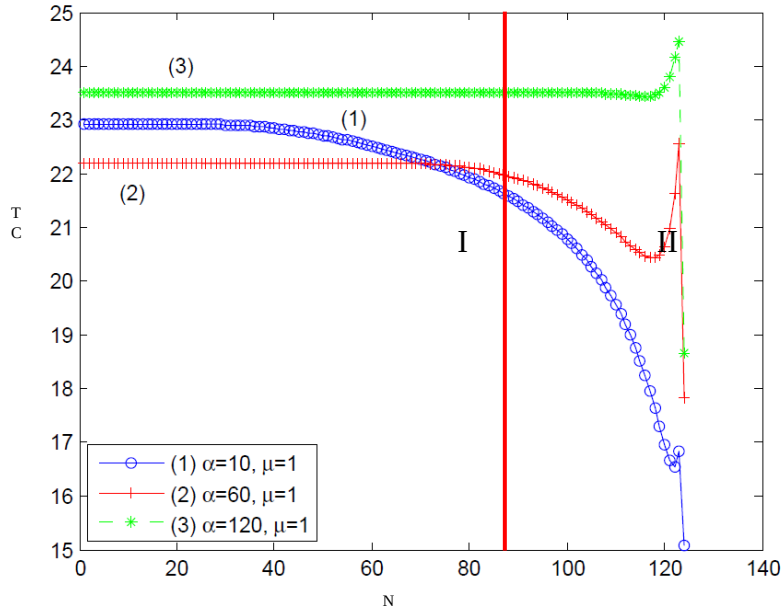

Figure 3. Threshold N versus managed cost $TC(N, 124)$ and μ = 1 during peak hours.

Fig. 3 demonstrates MC with a service rate of 1.5 (faster than the service rate of 1) during peak hours. The cost along with the reservation for important files (N is increasing) is gradually decreased to zero. This figure shows that because the service rate is fast, the cost will not increase. Further, the costs exhibit different shapes in areas I, II, III, and IV. In area I, the order of cost (MC_α) is $MC_{10} > MC_{60} > MC_{120}$. In area II, this order is $MC_{60} > MC_{10} > MC_{120}$. In area III, it is $MC_{60} > MC_{120} > MC_{10}$. Finally, in area IV, the order is $MC_{120} > MC_{60} > MC_{10}$. These different costs are attributed to the different values of C_h, C_r, and C_u. In area I, because of a small N, the blocking probability of the regular packets is large. Therefore, when the retrial rate is low, the retrial cost ($C_h L$) represents a major part of the cost; that is, when the retrial rate is low, the cost is large. However, in areas II and III, along with the increasing N, $C_r P_r$ and $C_u P_u$ represent the major parts of the total cost. $C_r P_r$ decreases more than $C_u P_u$ with an increase in N. Finally, in area IV, a larger retrial rate results in a higher value of MC.

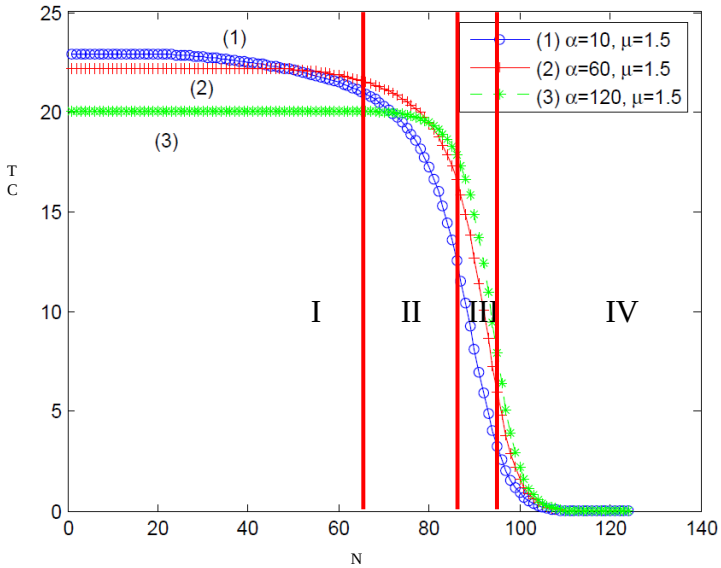

Figure 4. Threshold N versus managed cost $TC(N, 124)$ and μ = 1.5 during peak hours.

Fig. 4 shows MC with a service rate of 0.5 during regular hours. MC is smaller during regular hours than during peak hours. A large value of α implies a large value of MC.

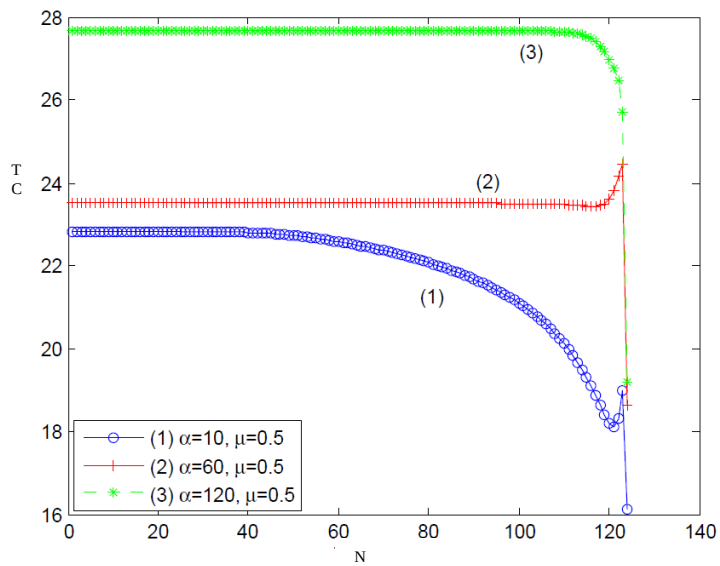

Figure 5. Threshold N versus managed cost $TC(N, 124)$ and μ = 0.5 during regular hours.

Fig. 5 illustrates MC with a service rate of 1 (faster than the service rate of 0.5) during regular hours. The cost along with the reservation for important files (N is increasing) is gradually decreased to zero. Further, along with different retrial rates (α), the costs change with different shapes: in area I, the order of cost (MC_α) is $MC_{10} > MC_{60} > MC_{120}$, but in areas II and III, the costs are $MC_{60} > MC_{10} > MC_{120}$ and $MC_{60} > MC_{120} > MC_{10}$, respectively. In area I, the increasing C_hL is larger than the increasing C_uP_u; that is, if the retrial rate is small, the cost is large. In areas II and III, along with the increasing value of N, because the decreasing rate of C_uP_u is smaller than the decreasing rate of C_hL, the costs change according to the difference in the decreasing rate.

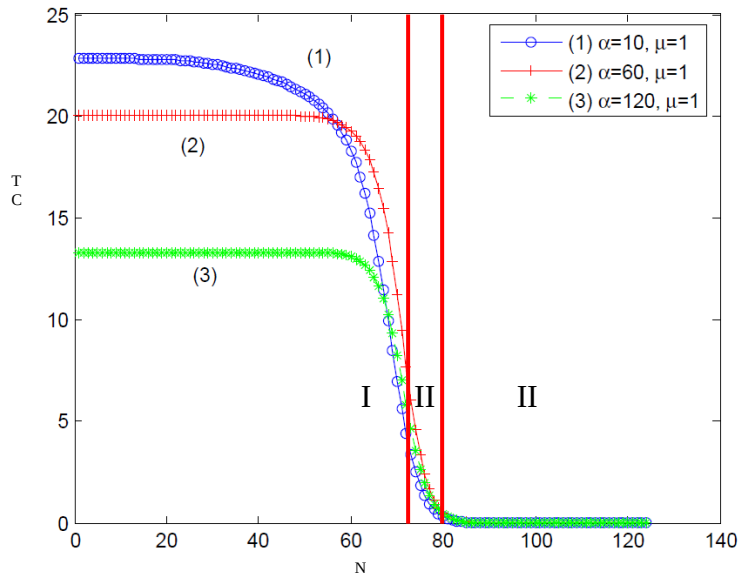

Figure 6. Threshold N versus managed cost $TC(N, 124)$ and μ = 1 during regular hours.

Fig. 6 shows MC with a service rate of 1.5 (faster than the service rate of 1) during regular hours. The changes in costs are similar to those shown in Fig. 5. Further, because the service rate in Fig. 6 is faster than that in Fig. 5, the cost in Fig. 6 decreases to zero at a faster rate than that in Fig. 6.

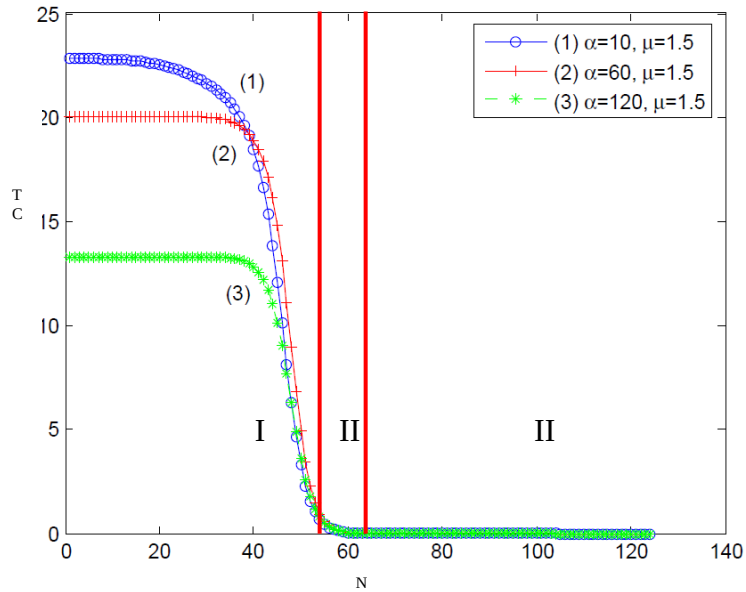

Figure 7. Threshold N versus managed cost $TC(M, 124)$ and μ = 1.5 during regular hours.

5. CONCLUSION

Previous studies have shown that a large-scale company can use 19 servers to reliably transfer the files of an important ordering system. However, along with the evolution of business, a content delivery service must be used carefully because of the sparse resources such as connections. Once a connection is selected for delivering content, that connection delivers no other packets. Therefore, it is important to discuss a way to use the connections efficiently. Further, each file to be delivered has its own specified importance. That is, each file has its own delivery cost. Therefore, the loss caused by the failure to deliver an important file is larger than that from the failure to deliver a regular file. Obviously, the costs of delivering important files and regular files must be treated as different. In order to use connections efficiently, we need to find the optimal cost. Therefore, in this study, we derived a cost function for content delivery through a queuing model.

By analyzing the queuing model, we found that during the peak hours, if the service rate is high (1.5), the cost will be gradually decreased to zero. Further, during regular hours, if the service rate is in the mid-range (1) or is high (1.5), the cost will be decreased to zero gradually. That is, the service rate is the key to providing better service. However, in order to increase the service rate, a company has to improve the performance of the servers and

increase the networking bandwidth. This will increase the cost considerably. In this study, we developed a way to find the lowest possible cost even when a company does not provide fast service. In summary, this work derived the content delivery cost function. Further studies are required to not only develop a method to measure the cost but also to derive a manageable cost function to illustrate an expanding business. Additionally, the reliability and validity must be verified and a comparison to other models must be considered in further studies.

ACKNOWLEDGEMENT

We especially thank Mr. Han-Hseng Huang for his contribution in the construction of the model.

REFERENCES

[1] W.M. Wang, C.C. Liang, H.Z. Lu, W.S. Chow, and K.Y. Chang, Research of testing process: the case of TOPS-system delivery process. *TL Technical Journal*, 34(1), p7-34, 2004.

[2] C. Ranganathan, and S. Ganpathy, Key dimensions of business-to-customer web sites. *Information and Management*, 39(6), p457-465, 2002.

[3] Y. Jiang, M.Y. Wu, and W. Shu, A hierarchical overlay multicast network. *Proceedings of IEEE conference on Multimedia and Expo*, June 30, Taipei, 2004.

[4] M. Fry, G. MacLarty, Policy-based content delivery: an active network approach. *Computer Communications*, 24(2), p241-248, 2001.

[5] C.C. Liang, C.H. Wang, H.S. Huang, and H. Luh, Measuring the cost of links of a disaster-avoided content-delivering service within a large-scaled company. *Proceedings of the 4th conference of Queueing Theory and Network Applications*, July 29-31, Singapore, 2009.

[6] C.C. Liang, C.H. Wang, H. Luh, P.Y. Hsu, and W. Yue, Proactive peer-to-peer traffic control when delivering large amounts of content within a large-scale organization. *Lecture Notes in Computer Science*, 5764(1), p217-228, 2009.

[7] T.N. Goh, A strategic assessment of six sigma. *Quality and Reliability Engineering International*, 18(5), p403-410, 2002.

[8] C.C. Liang, P.Y. Hsu, J.D. Leu, and H. Luh, An effective approach for content delivery in an evolving intranet environment - a case study of the largest telecom company in Taiwan. *Lecture Notes in Computer Science*, 3806(1), p740-749, 2005.

[9] C.C. Liang, C.H. Wang, H. Luh, and P.Y. Hsu, Disaster avoidance

mechanism for content-delivering service. *Computers and Operations Research*, 36(1), p27-39, 2009.

[10] H.Z. Lu, C.C. Liang, C.C. Chuan, and W.M. Wang, Discussion of TOPS/order software deployment, news publish and operation mechanism. *TL technical Journal*, 35(5), p719-733, 2005.

[11] C.C. Liang, C.R. Chuan, H.Z. Lu, and W.M. Wang, A software deploy model on TOPS/order system and its practice. *TL technical Journal*, 35(5-1), p19-27, 2005.

[12] R.J. Mills, D. Paper, K.A. Lawless, and J.M. Kulikowich, Hypertext navigation - an intrinsic component of the corporate intranet. *Journal of Computer and Information Systems*, 43(3), p44-50, 2002.

[13] M. Conti, E. Gregori, and W. Lapenna, Client-side content delivery policies in replicated web services: parallel access versus single server approach. *Performance Evaluation*, 59(23), p137-157, 2005.

[14] B.D. Choi, A. Melikov, and A. Velibekov, A simple numerical approximation of joint probabilities of calls in service and calls in the retrial group in a picocell. *Applied Computer Mathematics*, 7(1), p21-30, 2008.

[15] V.S. Korolyuk, and V.V. Korolyuk, *Stochastic models of systems*. Boston: Kluwer Academic Pluishers, 1999.

[16] Y.N. Lien, and Y.C. Wun, QoS-aware packet scheduling by looking ahead approach. *Proceedings of the 19th Workshop on Object-Oriented Technology and Applications*, September 26, Taipei, 2008.

[17] J. Xu, X. Tang, and W.C. Lee, Time-critical on-demand data broadcast: algorithms, analysis, and performance evaluation. *IEEE Transactions on Parallel and Distributed Systems*, 17(1), p3-14, 2006.

[18] M. Chen, X. Wang, R. Gunasekaran, H. Qi, and M. Shankar, Control-based real-time metadata matching for information dissemination. *Proceedings of the 2008 14th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, August 25-27, Kaohsiung, 2008.

[19] L.C. Wolf, L. Delgrossi, R. Steinmetz, S. Schaller, and H. Wittig, Issues of reserving resources in advance. *Lecture Notes in Computer Science*, 1018(1), 28-38, 1995.

[20] D. Ferrari, A. Gupta, and G. Ventre, Distributed advance reservation of real-time connections. *Multimedia Systems*, 5(3), p187-198, 1997.

[21] I. Foster, M. Fidler, A. Roy, V. Sander, L. Winkler, End-to-end quality of service for high-end applications. *Computer Communications*, 27(14) p1375-1388, 2004.