Measuring the Cost of Links of a Disaster-Avoided Content-Delivering Service within a Large-scaled Company

Liang, Chih-Chin
Dept. Business Administration, National Formosa University
No.64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan (R.O.C)
+886-9-12200671
lgcwow@gmail.com

Wang, Chia-Hung
Dept. Mathematical Sciences, National ChengChi University
NO.64, Sec.2, ZhiNan Rd., Wenshan District,Taipei City 11605,Taiwan (R.O.C)
+886-2-29393091 ext. 63961
93751502@nccu.edu.tw

Huang, Han-Sheng
Dept. Mathematical Sciences, National ChengChi University
NO.64, Sec.2, ZhiNan Rd., Wenshan District,Taipei City 11605,Taiwan (R.O.C)
+886-2-29393091 ext. 63961
jacky01140319@pchome.com.tw

Luh, Hsing
Dept. Mathematical Sciences, National ChengChi University
NO.64, Sec.2, ZhiNan Rd., Wenshan District,Taipei City 11605,Taiwan (R.O.C)
+886-2-29393091 ext. 67373
slu@nccu.edu.tw

ABSTRACT

A content-delivering service within a large-scaled company in Taiwan has been proposed to use 40 servers to meet user requirements on the previous study. The managers of the case company want to use resources as few as possible. From previous study [19-21], we suggest them using 19 servers to deliver large files. However, because we need not only servers but also links to deliver content from the source to all destinations, we cannot discuss servers merely. Therefore, we must discuss about the reliability and cost of links that in previous study it discussed few about this matter. Since the important packets must be delivered on time, and the regular packets are flexible. Therefore, we need a proper design to make the important packets deliver to the destinations without mistake. The regular packets also need to be delivered to destinations, but they can resent to destinations until the urgent files are delivered completely. Therefore, the reliability design of links must consider following utilization: the system is used to deliver every kind of data securely, but the important packets must be delivered without delay. Therefore, this work models the reliability and cost of links to find the proper resource allocation that can help manager delivering urgent files and regular files.

Keywords

Content delivery, Reliability

QTNA 2009, Jul 29-31, Singapore

1. INTRODUCTION

A large-scaled service sector regards uninterrupted customer service as one of the most important operational goals [12]. A typical large-scaled service company has various devices, such as personal computers, which are applied by clerks in operation centers to process business actions. Every device hosts numerous different client software applications, and each linking to the systems in the back offices. Generally, a company uses client-server or web-based architectures to deploy business systems [13,14]. Through client-server architecture, users can store applications and data, including promotion, area and customer codes, on the client devices. Through web-based architecture, users can utilize applications through web browser that a user cannot store applications on the client devices, but the user still needs to access data in the client machine. Obviously, no matter what architecture adopted, data will be exchanged among systems. For example, the large-scaled case company has ordering system and billing system that must exchange data to finish business operations. That is, with the interaction of architectures, web-based systems must write data to the devices for the references of client software of other client-server systems [15]. That is, to make a system consistent, we must deliver content, including data, information and updated files, through a reliable content-delivering service.

Additionally, because a large-scaled company has the hierarchical architecture, such a company requires a feasible approach to sending content along the hierarchy to gain support from all business units [16]. To ensure the reliability of content delivery, disaster avoidance is important to ensure that failed servers on the distribution routes do not bring the distribution to a halt [11]. Previous study describes the experience, which

originally applied one server to update the contents of all clerks' devices, and has since learned from the experience that disasters can stop the content-delivering service. Therefore, the company has to develop a new approach for delivering contents along a physical hierarchy. A content-delivering service called FnFDS (Fire and Forget Distribution System), which distributes the updated contents of the company, has been presented [11].

Previous studies also have analyzed that FnFDS based on the server number to find the reliability. The results show that FnFDS needs only 19 servers to serve content-delivering requirement and has outstanding performance: the content-delivering service will be failed 1.90735 occurrences per million hours. That is better than six-sigma capability: 3.4 occurrences per million hours [17-19]. This result only shows how many servers can handle file delivery, no matter how many links adopted to transfer packets to destinations [19]. However, the content-delivering service needs servers and connections to transfer packets. The reliability of the link is important to discuss. If we use links improperly, the important content might not be delivered to destinations.

With the evolvement of business environment the server can be utilized to disseminate every kind of information (files and messages) to more than 15,000 receivers, especially to the front-end clerks. In terms of resource usage, the connections between each server will be used frequently. Once the connections are used without considering cost efficiency, the content might not be delivered optimally; otherwise, the high-cost links will be used only when necessary [20-22]. That is, we must analyze the usage of connections to find out how to deliver content with efficiency.

Additionally, content has its specified privilege to be delivered. All kinds of content will be issued by operators, managers, and high-level managers. The content sent by high-level managers or the content is urgent to be delivered needs response immediately, but the regular packets sent by others are considered with lower priority. That is, the regular message can be delayed for delivery. Additionally, the company sometimes wants to deliver content immediately to all destinations even when the server load is high. That is, the urgent packets must be delivered on time. In summary, we must have the higher privilege to send urgent files. To understand the usage of connections, we must find the business requirements to content delivery service. Finally, we investigate a queueing model to analyze the usage of the links based on the scenario of the case company [23, 24].

The rest of this paper is structured as follows. Section 2 describes the usage of connections of FnFDS in details. Section 3 then describes the queueing model of links. Next, Section 4 discusses the content-delivering service based on the results of applying the queueing model. Finally, discussions and concluding remarks are presented in Section 5.

2. SYSTEM DESCRIPTION

To analyze the connections, we must understand the system architecture and its usage. The topology of FnFDS network is a completed network: each node has full links to other nodes within the same network.

Additionally, although FnFDS is on-line, the servers and links are used for not only file delivery but also every kind of packet delivery with the evolvement of business, including: mail, emergency information, daily updates, and instant messages. In other words, the network is utilized heavily. Furthermore, the packets can be delivered to not only all nodes but also partial nodes depend on the requirement. For example, once a message must be delivered to the northern business unit, the message won't be delivered to other business units. Additionally, each undelivered packets should be retrial to deliver to destinations immediately once connection error occurred. However, the blocking packets cannot be retrial forever. That is, the capacity of the sender is limited to let packets wait to be sent. Without loss of generality, we assume the queue for re-sending packets is limited. (Fig.1)

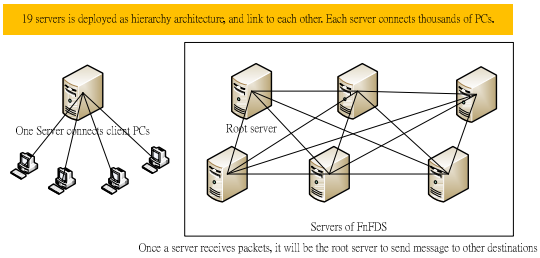

Figure 1. System Architecture

Using FnFDS, we can deliver packets through the root server mainly, but we still can use other servers to send files as well. That is, each server can be the role of root server. Therefore, once a server receives the sent file, the server will deliver the file to other linked servers.

From above description, in terms of cost efficiency, the manager must understand how to use the connections of a server (sender) efficiently. Because the important packets (urgent files) must be delivered on time, but the network might be occupied to transmit other packets (regular files). We must understand how many links must be used to send urgent files. Once there are K outbound lines of the root server, the cost analysis of this study is tried to find how many lines of K for transmitting urgent files. Namely, we must set up a signal for remaining lines to deliver urgent files. For example, if the signal is assigned to "N" connections, the connections will be reserved to deliver urgent packets only when "N" and above of connections are occupied.

Additionally, once the important message cannot be delivered, the business cannot go on properly. For example, once a notice of "system shutdown" cannot be sent to all receivers who need this information, the business operation should be terminated. Additionally, the frequency of sending packets on peak hours and regular hours is different. The peak hour means the managers send messages frequently, and the packets will be sent more often than regular hours.

The business operations are described as follows. In average, each file (packets) set 5.5 minutes to deliver to all destinations. The frequency of content delivery is set to transmit 5 times of general usage per minute and 2 times of important/emergency packet delivery per minute on the peak hour. The frequency of content delivery is set to transmit 0.06 times of general usage per minute and 0.01 times of important/emergency packet delivery per minute on the regular hour. The general usage packets should be retried for delivery after block by occupation of the connection.

That is, the packets enter a retrial group to wait the available network connections when they are blocked. The blocking packet in the retrial group will be sent after network is available to transfer packets. Additionally, the retrial group is finite because of the sparse business resources. Therefore, the message has the possibility to be discarded. From the experience, the probability of failing to re-send packets is almost 30% on the peak hour and 5% on the regular hour.

Finally, from the summary of the interview of managers, the cost of no delivering important packets is 20 times the cost of delivering packets successfully; the cost of no delivering regular packets is 2 times the cost of delivering packets successfully.

3. A QUEUEING MODEL

We investigate a queueing model for the connection usage of FnFDS in a large-scaled company. There are two types of content, one is the regular files and the other is urgent files. Let λ_r and λ_u be the average arrival rates of regular files and urgent files respectively. Based on the described scenario, we assume that there are K links (connections or transmission channels) in FnFDS, where $1 < K < \infty$. Link transmission rate (service rate) of both type of files (regular files and urgent files) is μ. If there exists at least one free link upon arrival of urgent files, it is admitted to be transmitted immediately. Otherwise, the attempt of transmission of urgent files is forcedly terminated. Let P_u be the probability that urgent files cannot be transmitted (do not receives service).

When there are N $(1 \le N \le K)$ or more links being busy, the incoming regular files will be blocked and enter the retrial group. The blocked regular files in the retrial group try to be re-sent after retrial time whose distribution is exponentially distributed with rate α. Here, we assume that capacity of the retrial group is finite, say R. This assumption is not extraordinary for intra-network content-delivering services. Let P_r be the probability that regular files cannot be transmitted (do not receives service) on its 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 blocked files in the retrial group is α. Let H_1 be the probability that regular files in the retrial group return to retrial group after an unsuccessful retrial. Then $1 - H_1$ is the probability that regular files in the retrial group leave the system after an unsuccessful retrial. The average number of regular files in retrial group is L.(Fig. 2)

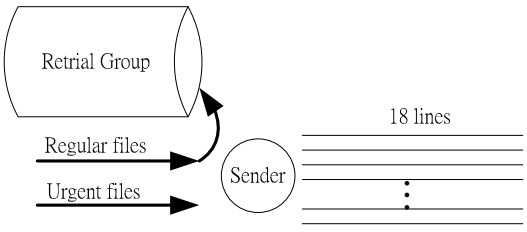

Figure 2. Queueing Model

We investigate the total expects cost function of this file transmission system. Our objective is to minimize the total expected cost by setting the threshold N given the number of links is K. We construct a total expected cost function (per unit time) for the N-signal $M/M/K$ queueing system, where N is the decision variable. The total expected cost function (per unit time) can be constructed as

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

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 files.

Let $X_1(t)$ be the random variable representing the number of blocked regular files in the retrial group at time t. Let $X_2(t)$ be the random variable representing 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

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

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

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

Solving the balance equation, we can obtain the stationary probability $P(i,j)$. Therefore, some important performance measures of this queueing system can be determined as follows. The average number of regular files in retrial group is

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

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

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

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

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

4. DISCUSSION

From the description in section 2, we find following terms:

1/μ = 5.5 min (each packet set should spend 5.5 minutes to deliver to all destinations),

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

K =18 (transmission channels),

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

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

$\lambda_r = 5$ (files/min) / $\lambda_u = 2$(files/min) at peak hour, and

$\lambda_r = 0.06$ (files/min) / $\lambda_u = 0.01$(files/min) at regular hour.

Additionally, to compare the differences of the frequency of retrying to send to destination, we assume following three parameters to find the differences:

$\alpha_1 = 12$ (retrying to send to destination per 5 seconds: the design of the FnFDS),

$\alpha_2 = 30$ (retrying to send to destination per 2 seconds), and

$\alpha_3 = 600$ (retrying to send to destination per 0.01 seconds).

We have following results. We found that at the peak hour, the smaller the frequency of retrying to send packets to destination (α is larger) is, the larger the total expected cost (shown in Fig. 3). Additionally, along with the increasing N, the total expected cost is fixed until the point N ($N = 15$ on $\alpha_1 = 12$; $N = 16$ on $\alpha_2 = 30$; $N = 17$ on $\alpha_3 = 600$). When $N$is larger than the rising point, the expected cost will increase along with the increasing N. Fig. 4 shows that at the regular hour, no matter what α is, the total expected cost has almost no difference along with the increasing threshold N. It shows that the optimized N (that can meet the minimum cost) is increased along with the increasing α.

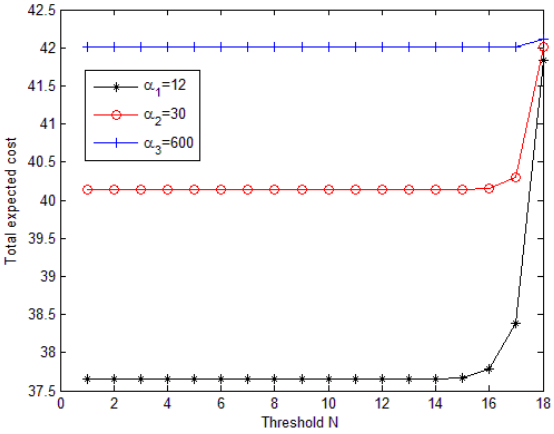

Figure 3. The threshold N versus total expected cost $TC(N, 18)$ at peak hour.

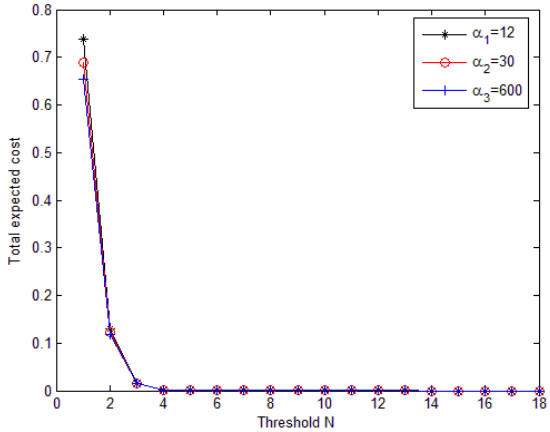

Figure 4. The threshold N versus total expected cost $TC(N, 18)$ at regular hour.

Fig. 5 shows that at the peak hour, along with the increasing threshold N, the smaller the frequency of retrying to send packets to destination is, the smaller the average number of regular files in the retrial group is fixed until the point N($N = 15$ on $\alpha_1 = 12$; $N = 16$ on $\alpha_2 = 30$; $N = 17$ on $\alpha_3 = 600$). At the regular hour, no matter what α is, we can expect the same results as at regular hour: the average number of regular files in the retrial group is down very soon along with the increasing threshold N (shown in Fig. 6).

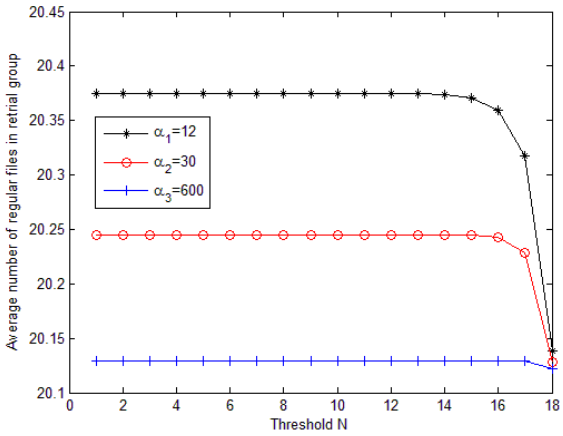

Figure 5. The threshold N versus average number of regular files in retrial group L at peak hour.

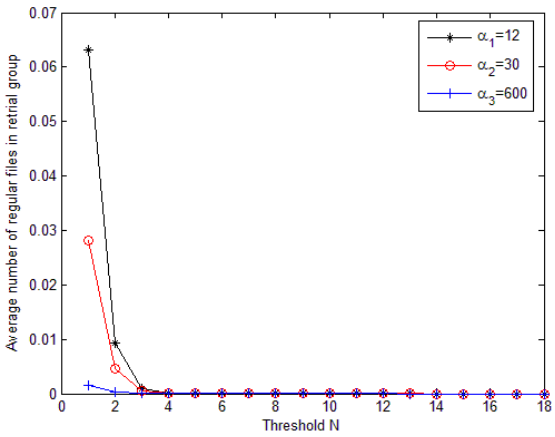

Figure 6. The threshold N versus average number of regular files in retrial group L at regular hour.

At the peak hour, along with the increasing threshold N, no matter what α is, those curves of probabilities that regular files cannot be transmitted on its first attempt files are decreasing and concave until the point N($N = 15$ w $\alpha_1 = 12$; $N = 16$ on $\alpha_2 = 30$; $N = 17$ on $\alpha_3 = 600$) (shown in Fig. 7). Fig. 8 shows that at the regular hour, no matter what α is, those curves of probabilities that regular files cannot be transmitted on its first attempt files are decreasing and convex. We find that the probability at regular hour is decreased faster than that at peak hour along with the increasing threshold N.

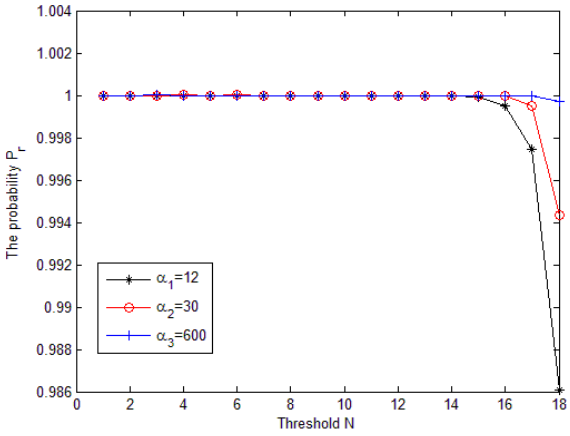

Figure 7. The threshold N versus the probability that regular files cannot be transmitted on its first attempt P_r at peak hour.

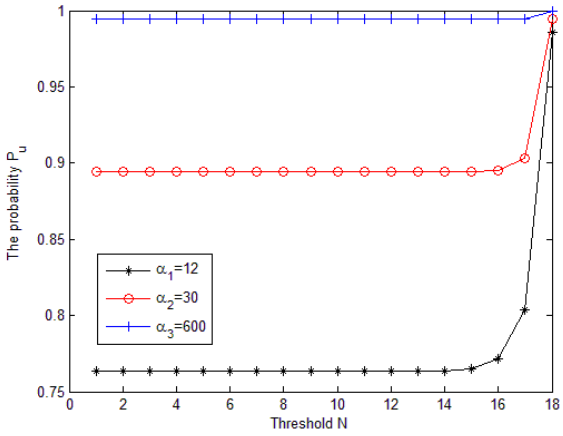

Figure 9. The threshold N versus the probability that urgent files cannot be transmitted P_u at peak hour.

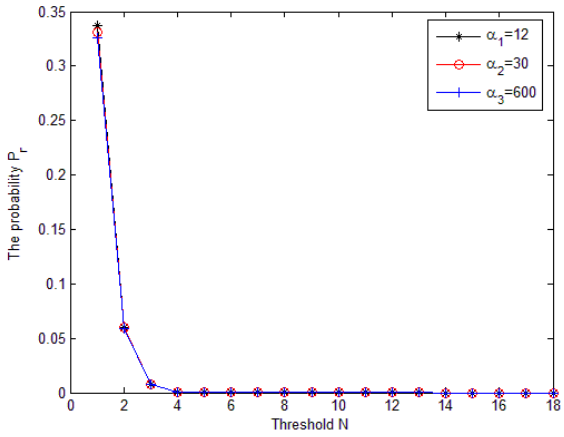

Figure 8. The threshold N versus the probability that regular files cannot be transmitted on its first attempt P_r at regular hour.

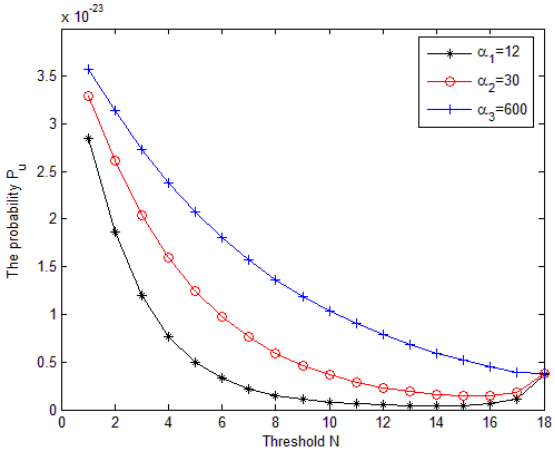

Figure 10. The threshold N versus the probability that urgent files cannot be transmitted P_u at regular hour.

Finally, along with the increasing threshold N, at the peak hour or at the regular, no matter what α is, those curves of probabilities that urgent files cannot be transmitted on its first attempt files are convex until the point N ($N = 15$ w $\alpha_1 = 12$; $N = 16$ on $\alpha_2 = 30$; $N = 17$ on $\alpha_3 = 600$) (shown in Fig. 9 and Fig. 10). The probability at peak hour is increased early than at regular hour.

5. CONCLUSION

Along with the business evolvement, the sparse resources must be used carefully, especially for a large-scaled company. Previous studies show that a large-scaled case company employees a content delivery application to transfer files of the important back-office system. To transfer packets, we need servers and connections (links). Previous studies also show that 19 is an optimized number of servers to help the application to deliver content. However, the connection of the application has not been discussed. Because of the sparse networking resources, we must understand the cost of links to help the manager make decision to adopt proper networking facilities to build up the content delivery service. Therefore, in this work, based on the scenario provided by the case company, we try to analyze the cost of links.

Additionally, each delivering file has its importance. Once an important file needed to be sent, the connections must be available to transmit the file. For example, if there are K outbound lines of the server, the cost analysis of this study is tried to find how many

lines (N) of K we should use for transmit urgent files. That is, we must find out how to set up a proper signal for remaining lines to deliver urgent packets.

Through analyzing the queneing model, we found that at the peak hour, the smaller the frequency of retrying to send packets to destination (α is larger) is, the larger the total expected cost presents along with the increasing threshold N. Additionally, along with the increasing N, the total expected cost is decreasing firstly, and then meeting the minimum cost, finally increasing along with the increasing N. At the regular hour, no matter what α is, the total expected cost has almost no difference along with the increasing threshold N.

Finally, through our proposed scenarios and assumptions, we found the total expected costs and the optimized N will be found (in terms of the minimum point): N =17 (when $\alpha = 600$), N =16 (when $\alpha = 30$), and N =15 (when $\alpha = 12$). That is, once the other parameters are fixed, the optimized N (that can meet the minimum cost) is decreased along with the frequently resending packets to the destination.

ACKNOWLEDGMENTS

This work was supported in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 19500070) and MEXT.ORC (2004-2008), Japan.

7. REFERENCES

[1] King, D. 2002. Post Disaster Surveys: experience and methodology. Australian J. Emerge. Mana. 17,3, 1-13.

[2] Pidgeon, N. 2000. O'Leary M. Man-made disasters: why technology and organizations (sometimes) fail. Safe. Sci. 34, 15-30.

[3] Hayes, P.E. Hammons A. 2002. Picking up the pieces: Utilizing disaster recovery project management to improve readiness and response time. IEEE Ind. Appl. Mag. 8, 6, 27-36.

[4] Bank, D.R. Telecomm disaster recovery planning for electric utilities. Proceedings of IEEE conference on Rural Electric Power (2005). REP '05. IEEE Press.

[5] Shao, B.B.M. 2005. Optimal redundancy allocation for information technology disaster recovery in the network economy. IEEE Trans. Depend. Sec. Comput. 2, 3, 262-267.

[6] Fallara, P. 2003. Disaster recovery planning. IEEE Potentials. 22, 3, 42-44.

[7] Smith, D.R. Cybrowski, W.J., Zawislan, F., Amstein, D., Dayton, A.D., and Studwell, T.D. 1994. Contigency/Disaster recovery planning for transmission systems of the defense information system network. IEEE Select. Areas Commun. 12, 1, 13-22.

[8] Hiles, A. 1992. Surviving a computer disaster. Eng. Manage. 3, 3, 271-274.

[9] Stuckenschmidt, H., van Harmelen, F. 2004. Generating and managing metadata for web-based information systems. Knowledge-Based Syst. 17, 5-6, 201-206.

[10] Taylor, M.J., Mcwilliam, J., England, D., Akomode, J. 2004. Skills required in developing electronic commerce for small and medium enterprises: case based generalization approach. Elec. Commerce Res. & Appli. 3, 3, 253-265.

[11] Fry, M., MacLarty, G. 2001. Policy-based content delivery: an active network approach. Comput. Commun. 24, 241-248.

[12] Wang, W.M., Liang, C.C., Lu, H.Z., Chow, W.S., Chang, K.Y. 2004. Research of Testing Process: The Case of TOPS-System Delivery Process. TL. Tech. J. 34, 1, 7-34.

[13] Mills, R.J., Paper, D., Lawless, K.A., Kulikowich, J.M. 2002. Hypertext navigation- an intrinsic component of the corporate intranet. J. Comput. Inform. Syst. 43, 3, 44-50.

[14] Conti, M., Gregori, E., Lapenna, W. 2005. Client-side content delivery policies in replicated web services: parallel access versus single server approach. Perform. Eva. 59, 23, 137-157.

[15] Ranganathan, C., Ganpathy, S. 2002. Key dimensions of business-to-customer web sites. Inform. & Mana. 39, 457-465.

[16] Jiang, Y., Wu, M.Y., Shu, W. 2004. A hierarchical overlay multicast network. Proceedings of the IEEE conference on Multimedia and Expo. ME '04. 2004: 1047-1050.

[17] Goh, T.N. 2002. A strategic assessment of six sigma. Qual. Reliab. Eng. Int. 18, 5, 403-410.

[18] Liang, C.C., Hsu, P.Y., Leu, J.D., Luh, H. 2005. An effective approach for content delivery in an evolving intranet environment- a case study of the largest telecom company in Taiwan. Lect. Notes. Comp. Sci. 3806, 740-7 49.

[19] Liang, C.C, Wang, C.H., Luh, H., Hsu, P.Y. 2009. Disaster Avoidance Mechanism for Content-Delivering Service. Comput. Oper. Res. 36, 1, 27-39.

[20] Lu, H. Z., Liang, C. C., Chuan, C. C., and Wang, W. M. 2005. Discussion of TOPS/order software deployment, news publish and operation mechanism. TL. tech. J. 35, 5, 719-733.

[21] Liang, C. C., Chuan, C. R., Lu, H. Z., and Wang, W. M. 2005. A software deploy model on TOPS/order system and its practice. TL. tech. J. 35, 5-1, 19-27.

[22] Choi, B.D., Melikov, A., Velibekov, A. 2008. A simple numerical approximation of joint probabilities of calls in service and calls in the retrial group in a picocell. Appl. Comput. Math. 7, 1, 21-30.

[23] Korolyuk, V.S., Korolyuk, V.V. 1999. Stochastic models of systems. Kluwer Academic Pluishers, Boston.