Optimization of storage quota based on user's usage distribution

Kasahara, Yoshiaki
Research Institute for Information Technology, Kyushu University

Kawatani, Takuya
Graduate School of Information Science and Electrical Engineering, Kyushu Uiversity

Ito, Eisuke
Research Institute for Information Technology, Kyushu University

Simozono, Koichi
Computing and Communications Center, Kagoshima University

他

出版情報：Proceeding of the 2015 IEEE 39th Annual Computer Software and Applications Conference Workshops (COMPSACW 2015), pp.149-155, 2015-07-05. IEEE
バージョン：
権利関係：

Optimization of storage quota based on user's usage distribution

Yoshiaki Kasahara*, Takuya Kawatani†, Eisuke Ito*, Koichi Simozono‡ and Naomi Fujimura§
*Research Institute for IT, Kyushu University, Fukuoka 812-8581, Japan
yoshiaki.kasahara.820, ito.eisuke.523@m.kyushu-u.ac.jp
†Department of ISEE, Kyushu University, Fukuoka 812-8581, Japan
kawatani.takuya.827@s.kyushu-u.ac.jp
‡Computing and Communications Center, Kagoshima University, Kagoshima 890-0065, Japan
simozono@cc.kagoshima-u.ac.jp
§Department of Design, Kyushu University, Fukuoka 812-8581, Japan
fujimura.naomi.274@m.kyushu-u.ac.jp

***Abstract*—To prevent shortage of storage space in a service system, an administrator usually set per-user quota as an upper limit of usable space for each user. To avoid service failure caused by resource exhaustion, the administrator tends to set a conservative quota value such as the storage capacity divided by the expected maximum number of users. In this research, we analyzed long-term storage usage history of our email system and file sharing system in Kyushu University. Mostly through the analyzed period, the usage pattern showed a long-tailed distribution similar to log-normal distribution. Also the overall storage consumption slowly increased during the analyzed period. Based on these analysis, we defined "storage utilization ratio" to evaluate how the storage was effectively used. By approximating a storage utilization pattern as a power-law distribution, we proposed a method to calculate the optimal quota value to maximize the utilization ratio.**

I. Introduction

The storage capacity of a system is usually limited. To prevent shortage of storage space in a service system, the administrator of the service system usually set user's quota as an upper limit of usable space for each user. To avoid service failure caused by resource exhaustion, the administrator tends to set a conservative quota value such as the total storage capacity divided by the expected maximum number of users. As we discuss in this paper later, the actual usage pattern exhibited a long-tailed distribution similar to log-normal distribution. Only a fraction of users fully used up their quota limit, some users consumed up to 60% of their quota, and majority of remaining users only consumed less than 20%. In consequence of such a long-tailed usage pattern, applying a conservative quota value causes low utilization of the overall storage resource.

In this research, we analyzed users' storage consumption of some services in order to estimate the optimal quota value which improves the utilization of the storage capacity. Specifically, we analyzed the storage usage history of a university-wide email system (called "Primary Mail Service") and file sharing system in Kyushu University. After that, we discussed how to estimate the optimal quota value setting which improv the utilization efficiency of storage resource based on per-user storage consumption distribution. By modeling the distribution, we could estimate the optimal quota value from the number of users and the whole storage capacity.

The rest of this paper is organized as follows. In section II, we describe some related studies. In section III, we introduce our email system and file sharing system in Kyushu University which are the target of our analysis. In section IV we analyze the distribution of per-user storage usage in these systems. In section V, we discuss how to estimate the optimal quota value based on the analysis in section IV. Finally, we present our conclusion and future works in section VI.

II. Related work

In article [1], Mitzenmacher mentioned that file size distributions were best modelled by a power-law distribution or a log-normal distribution. He surveyed about log-normal distribution and power-law distribution and reported that these distributions appeared frequently in various phenomena studied in economics and natural science. Also he showed that log-normal distributions had arisen as a possible alternative to power law. In our work, we also confirmed that user's storage usage distribution matched well with log-normal distribution in mail system and file sharing system of Kyushu University. As Mitzenmacher described in [1], we also didn't know an explicit model why the distribution was similar to log-normal distribution. It is our future work to establish a user behavioral model to explain how the distribution is formed.

In article [2], Kuninaka, et al. mentioned that a log-normal distribution appeared in various complex systems. They showed that many phenomena considered to be a normal distribution were actually fit better to a log-normal distribution. For example, people's height distribution was considered to be a normal distribution, but children's height distribution in a growth period fit better to a log-normal distribution. Also they showed that log-normal was more appropriate than power law or Zipf's law for representing population of cities [3]. User's storage usage could be considered to be in a growth period because it grows in time course, and might form a log-normal distribution. We want to study that more closely later.

In article [4], McKnight discussed future planning of storage preparation in information systems. He showed that "(total volume size) = (number of users) * (user quota) * 1.4" is

TABLE I. THE NUMBER OF IDS IN KYUSHU UNIV.(JAN. 2015)

Role	Total No. of IDs (approx.)
Curricular students	19,000
Non-curricular students	500
Faculty and staff members	9,000
Non-employee workers	1,000
Total	30,500

a reasonable starting point for the initial purchase. It was based on his professional experience as a data storage administrator, and the reasoning was not discussed.

III. TARGET SYSTEMS

In this research, we analyzed the university-wide email system and file sharing system in Kyushu University. In this section, first we briefly introduce the number of staff members and students in Kyushu University, which is the number of users for these systems. Next, we describe the details of the systems.

A. Number of Users in Kyushu University

TABLE I shows the approximate number of IDs issued by the university-wide authentication service[5] in Kyushu University as of January 2015. The number also represents the number of users of the mail system.

B. Kyushu University Primary Mail Service

The mail service analyzed in this study was called "Kyushu University Primary Mail Service"[6]. It was provided as one of the most important communication infrastructure in Kyushu University. The system we analyzed was for staff members, and operated from July 2009 to March 2014 and then replaced with newer system. There was another system for students, but we didn't analyze it due to lack of usage history.

Table II shows some numbers related to the resource and limitation of the system from July 2009 to March 2014. The column of "Date" denotes when the values of the row had become in effect. The system provided SMTP, POP, and webmail for free. From January 2011, a premium service class for paid users was started. The service included 10GB quota without message expiration and IMAP support. In April 2013, the quota value was expanded to 20GB due to requests from paid users.

Furthermore, the quota value was actually a "soft" limit, which meant that a user could store messages even after the using storage size exceeded the quota value. The system sent a warning message to the mailbox of such a user, but didn't block further incoming message. It was because there was a retention period for messages of non-paid users, and expired messages were automatically removed from the system.

TABLE II. RESOURCE AND LIMITATION (PRIMARY MAIL SERVICE)

Date	Jul.2009	Dec.2009	Dec.2011
Total storage	1,200 GB	1,200 GB	2,600 GB
User quota	100 MB	100 MB	300 MB
Expire	30 days	60 days	60 days
Message size	Max 20MB/message		

TABLE III. RESOURCE AND LIMITATION (FILE SHARING SYSTEM)

Date	Aug.2010	Jan.2011	Jun.2011	Sep.2011
Total storage	700GB			
User quota	1GB			
Expire	14 days	28 days		90 days
Message size	Staff members		Staff+Students	

C. File Sharing System

From August 2010, Kyushu University have been providing a service called "File Sharing System" targeted for all the staff members. The system was built with commercial software for building online storage system named "Proself" developed by Northgrid.

Originally the service was designed in order to reduce the size of email messages by storing large files and providing URL for the files instead of attaching them. Therefore the system was configured to delete old files after a retention period had passed. TABLE III shows actual limitation values. The column of "Date" denotes when the values of the row had become in effect.

The service extended the retention period a couple of times based on many requests from users. Also from June 2011, the service was also available for students. There were strong demands from paid users of Primary Mail Service to expand the user quota of File Sharing Service. In response to the demands, the user quota of paid user was expanded to 10GB in September 2011. It was expanded again to 20GB in April 2013.

In contrast with Primary Mail Service, the quota value of File Sharing System was "hard" limit. A user couldn't store files anymore if the total amount of file size exceeds his/her allowed capacity.

IV. STORAGE USAGE DISTRIBUTION ANALYSIS

Primary Mail Service had recorded storage usage per user from May 2009 to March 2014 (until the end of the service) every day. On the other hand, File Sharing System didn't have such a long-term record, and we obtained a snapshot of usage status manually on January 27th, 2014. We used these data to analyze the number of active users and the situation of storage usage.

A. Storage usage distribution of File Sharing System

We sorted each user's amount of storage usage in descending order to determine the usage rank of each user. Fig. 1 shows a log-log scatter plot of rank vs size (each user's storage usage) distribution of File Sharing System on January 27th, 2014. The vertical axis is usage amount in MB, and the horizontal axis is the rank by the usage amount.

The distribution of the storage usage exhibited like a long tail form. Only a fraction of users used large volume, and most users only consumed small capacity.

Next, we estimated the distribution of storage usage by a nonlinear regression analysis using data on Fig. 1. Top rank users and bottom rank users were excluded as follows. The lower rank users held no user's file stored in the system, so these were excluded. Higher rank users exceeding 1GB

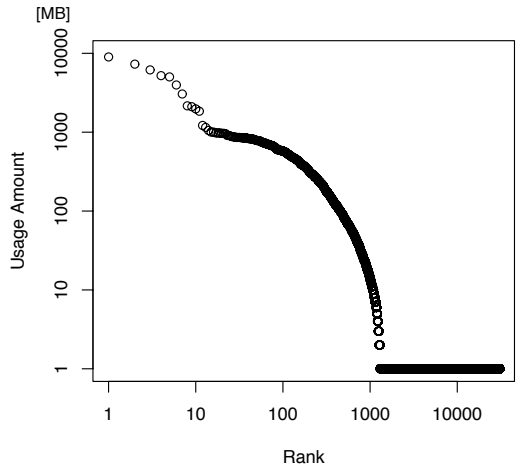

Fig. 1. File Sharing System: Rank-usage (log-log) (Jan.27th, 2014).

TABLE IV. AIC (FILE SHARING SYSTEM)

Distribution	AIC
Power law	15966.28
Exponential	10559.04
Log-normal	9244.65

quota were actually paid users. They were also excluded because their quota value was 20GB and considered outliers. In summary, we included 1,287 users whose usage amount was less than 1GB and more than 2MB for the estimation. In Fig.1, these users were from rank 15 to 1,302, and the partial plot was like an arched curve.

We selected three kinds of long-tail distribution, a power-law distribution, an exponential distribution, and a log-normal distribution as candidates for this storage usage distribution, and nonlinear regression analyses were performed against each of them. We used R [7] for nonlinear regression analyses. To evaluate how each candidate of distribution model fit with the actual distribution, we calculate the value of AIC (Akaike's Information Criterion). Calculated AIC values are shown in TABLE IV.

A model with the smallest AIC value is the optimal model. According to TABLE IV, a log-normal distribution was optimal among three. Fig. 2 shows the calculated values of the estimated log-normal distribution model plotted with the actual distribution. The two curves are close to each other except several highest rank users. Please note that this graph is in log-log scale, so the actual difference of lower rank users are small compared to higher rank users.

B. Storage usage distribution of Primary Mail Service

Primary Mail Service had recorded storage usage per user every day. We sorted each user's amount of storage usage in descending order and plotted a log-log scale graph with the vertical axis of usage amount and the horizontal axis of the user rank. For example, Fig. 3 shows the distribution on January 27th, 2014.

The red horizontal line in Fig.3 indicates the quota value (300MB) of non-paid (ordinary) users. The blue vertical line

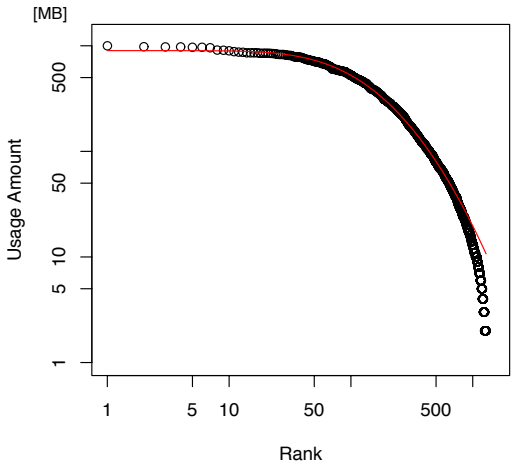

Fig. 2. File Sharing System: Actual data (dots) and model data (line).

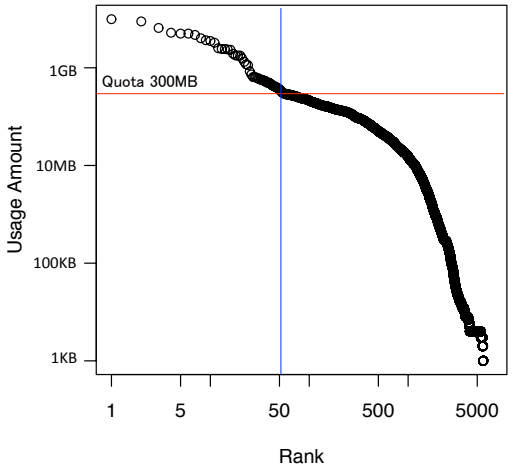

Fig. 3. Primary Mail Service: rank-usage (log-log) (Jan.27th, 2014).

indicates the rank (54th) where the user's storage usage exceeded the quota value. The user rank 1st to 54th consisted of two user groups. All the users from 1st to 24th (the green vertical line) were paid users whose quota value was 10GB, and actually used more than 300MB. Users from 25th to 54th were non-paid (ordinary) user except one paid user, but they exceeded their quota value of 300MB. As described before, it was because the quota value was a "soft" limit. The system sent a warning message to the mailbox of such a user, but didn't block further incoming message nor delete old messages automatically.

We analyzed the distribution of storage usage by a non-linear regression analysis like section IV-A. Users using more than 300MB and inactive users who didn't use storage space were excluded as outliers. In summary, we included 4,139 users whose usage amount was less than 300MB and more than 5KB for the estimation. In Fig. 3, these users were in the right side of the blue vertical line. Calculated AIC values after the regression analyses are shown in TABLE V.

According to TABLE V, again a log-normal distribution was optimal among them for estimating the usage distribution of Primary Mail Service. Fig. 4 shows the calculated values

TABLE V. AIC (PRIMARY MAIL SERVICE)

Distribution	AIC
Power law	95892.74
Exponential	81720.74
Normal	98252.47
Log-normal	79958.36

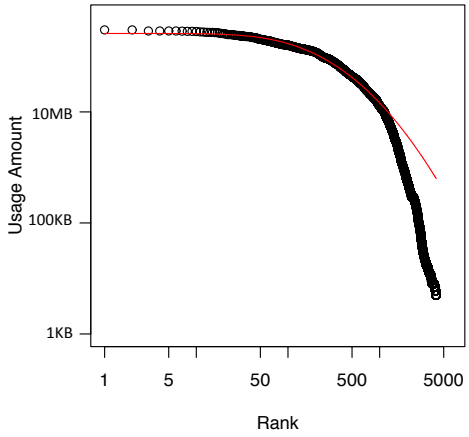

Fig. 4. Primary Mail Service: Actual data (dots) and model data (line).

of the estimated log-normal distribution model plotted with the actual distribution (usage amount is in KB). In this case, the usage amount of top 2,000 users almost agreed with each other, but lower rank users was over-estimated.

C. Long term trend of storage utilization

We investigated a long-term transition of storage utilization in Primary Mail Service. We used 15th of every month from February 2011 to February 2014 for the analysis, and the result is shown in Fig. 5. The black line indicates the total amount of storage usage, followed by top 10% (in red), top 5% (in yellow), and top 1% (in blue) of users consumed. Paid users and inactive users were excluded as outliers.

Fig.5 shows that top 10% users consumed more than 90% of total storage in use. Throughout the operation period of the system, most users used little storage space, and only a few heavy users consumed much storage space.

V. ESTIMATING OPTIMAL QUOTA VALUE

In this section, we discuss about optimal quota value. Users would like to use the capacity as much as possible, and want no restriction. On the other hand, the real storage capacity is limited, so a system administrator would like to suppress usage within the actual capacity. Also there should be some margin to cope with sudden increase of incoming data or gradual increase with time. Therefore an administrator always want to reserve a certain amount of space.

A. Symbols for modeling

We use the following symbols for modeling.

S : Upper bound of storage usage,
u_i : Storage usage amount for user i,
U : Total usage amount ($U = \sum u_i$),
q : Quota for users ($\forall i, u_i \leq q$) .

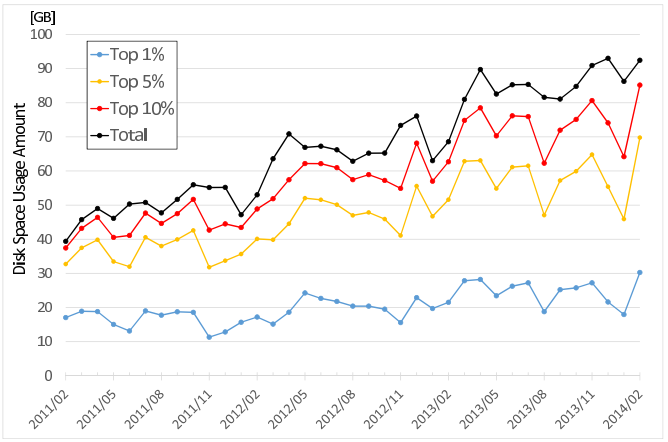

Fig. 5. Long-term transition of storage usage (Feb.2011–Feb.2014)

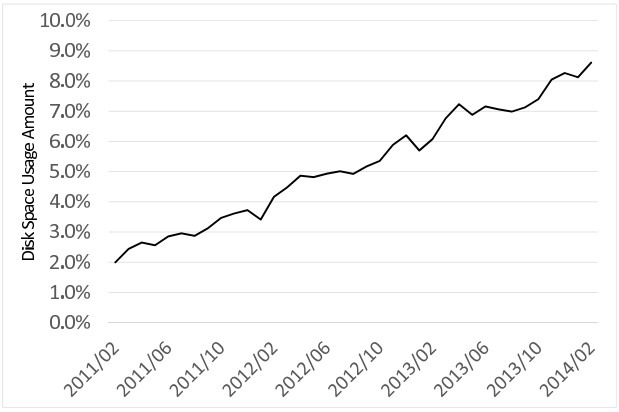

Fig. 6. Distribution change after increasing the quota value

S denotes the usable storage space decided by an administrator, and it is not the total amount of storage in the system. For example, there is storage of capacity 1TB in a mail system, and the administrator thinks users may use up to 75% of the total capacity, then S is set to 750GB.

We defined U/S as "storage utilization ratio" to evaluate how the storage was effectively used.

B. Storage utilization ratio of Primary Mail Service

Fig. 6 shows the storage utilization ratio of Primary Mail Service from February 2011 to February 2014. We set the upper bound of storage usage S as 2,000GB (=75% of the total storage capacity). The utilization ratio was very low throughout the time period. At that time, administrators had just considered the worst case scenario that all the users would use storage up to their assigned quota and set the quota value very small.

C. Approximation by Power Law Distribution

The actual storage consumption among users wasn't homogeneous as shown in section IV. It was like a long-tail distribution such that only a fraction of users used up to their

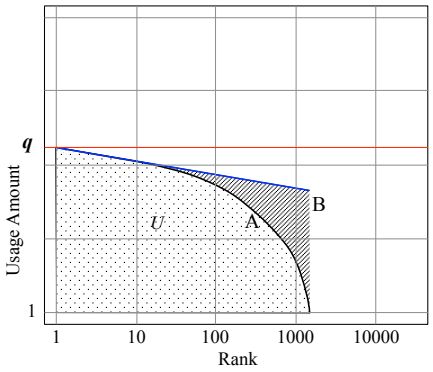

Fig. 7. The real distribution and its approximation by a power-law distribution

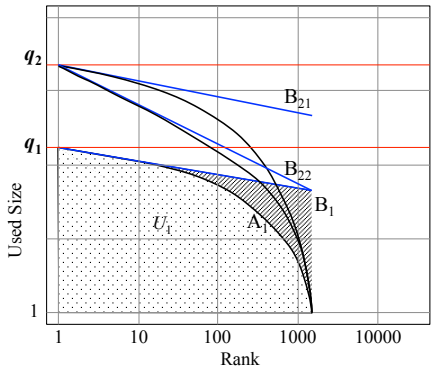

Fig. 8. Distribution change after increasing the quota value

quota limit and most of users only used little space. Based on the distribution, we estimated a quota value to increase the storage utilization ratio.

First we approximated the distribution as a power-law distribution. As shown in section IV, the most suitable distribution was a log-normal distribution, but we considered that a power-law distribution is more suitable to estimate the optimal quota value because we could simplify calculation. It is known that higher rank part of a log-normal distribution can be approximated by a power-law distribution.

A power-law distribution is represented by (1).

$$y = f(x) = qx^a, \tag{1}$$

where a is a scaling exponent. To take logarithm of both side of (1), it is transformed as follows:

$$\log y = \log qx^a = a \log x + \log q. \tag{2}$$

Let $Y = \log y, X = \log x, Q = \log q$, then we obtain

$$Y = aX + Q. \tag{3}$$

Equation (3) is a liner function with gradient a and intercept Q, so we could simplify the estimation.

Fig.7 shows the relationship between the real distribution and approximation in a log-log plot. The vertical axis is usage amount and the horizontal axis is the rank by the usage amount. We assume that the curve A denotes the actual usage distribution, and the line B denotes approximated values by a power-law distribution.

The area under the curve A is the actual storage usage U, and the area between A and B is the error of estimation. We could make sure that B was always above A, then the error became an extra capacity margin. Also please note that the graph is in log-log scale, so the error is relatively small.

D. Estimation of the optimal quota value

Consider changing the quota value from q_1 to q_2 $(q_1 < q_2)$ in order to increase the storage utilization ratio. Let U_2 as the storage usage when the quota value was changed to q_2. We need to select q_2 where U_2 won't exceed the total usable capacity S while increasing the storage utilization ratio.

Let us discuss how the usage distribution would change after changing the quota value using Fig. 8. After changing the quota value from q_1 to q_2, users' usage distribution would form new log-normal distribution.

B_1 denotes an approximated distribution under the quota value q_1. If new distribution retains similar gradient under the quota value q_2, the distribution will be similar to B_{21} in Fig.8. There is another possibility that it will become like B_{22} with steeper gradient, but currently we don't have enough clue which is correct. It is our future work to observe the real system to see how the usage distribution will change after the quota value is increased.

To change the quota value in order to increase the storage utilization ratio, it is better to use an estimation which overestimate users' consumption. In Fig. 8, B_{21} consumes more storage space than B_{22}, so we consider that B_{21} is more appropriate for the purpose.

By using estimation by a power-law distribution (1), the total usage amount U is as follows:

$$U = \sum_{x=1}^{n} f(x) = q \sum_{x=1}^{n} x^a, \tag{4}$$

where n is the number of users in the system. $U = S$ when the storage utilization ratio U/S is 100%. In that case, the quota value q is

$$q = \frac{S}{\sum_{x=1}^{n} x^a}. \tag{5}$$

E. Optimal quota value for Primary Mail Service

As shown in subsection V-B, the storage utilization ratio of Kyushu University Primary Mail Service was very low at that time. In this section, we estimated the quota value to make the utilization ratio as 100%. We used the usage distribution as of January 27th, 2014. To simplify, we excluded paid users, users whose usage was exceeded their quota value, and inactive users who stored nothing in the system. The number of remaining ordinary users was 4,139.

We used R for nonlinear regression analyses to estimate the gradient of the power-law distribution B in Fig. 7. We used only the top 400 users for the estimation, because adding lower rank users will make the gradient unnecessarily steeper. At the result, we got the gradient (scaling exponent) as $a = -0.404485$.

On January 27th, 2014, the storage usage of ordinary users we took into consideration was 70GB. We calculated the value of U by (4) with the gradient -0.40 and we got U = 71.91GB.

The total storage capacity of Primary Mail Service was 2,600GB at that time. We assumed the upper bound S as 2,000GB (= 75% of the total), and we could estimate the optimal quota value as about 8.1GB by (5).

F. Optimal quota value for File Sharing System

The storage utilization ratio of file sharing system was also low at that time. In this subsection, we estimated the quota value to make the utilization ratio as 100%. As same as the mail system, we used the usage distribution as of January 27th 2014, and we excluded paid users for simplify. The number of remaining ordinary users was 1,287.

We used R for nonlinear regression analyses to estimate the gradient of the power-law distribution B in Fig. 7. At the result, we got the gradient (scaling exponent) as $a = -0.4516$.

The total storage capacity of File Sharing System was 700GB at that time. We assumed the upper bound S as 525GB (= 75% of the total), and we could estimate the optimal quota value as about 5.75GB by (5).

VI. Conclusion

In this research, we analyzed long term storage usage history of our email system and file sharing system in Kyushu University. The storage usage distribution exhibited a long-tailed distribution. Nonlinear regression analyses showed that it was similar to a log-normal distribution. Also observing long-term history showed that the overall storage consumption slowly increased during the analyzed period.

To improve the storage utilization efficiency, we defined "storage utilization ratio" and proposed a method to evaluate how the storage was effectively used. By approximating a storage utilization pattern as a power-law distribution, we proposed a method to calculate the optimal quota value to maximize the utilization ratio. Based on actual data, we calculated the optimal quota value to maximize the utilization ratio of our email service at that time.

As a future work, we will apply the method we proposed to our system we are operating now. We are collecting usage history of new email system (started in March 2014)[8] and file sharing system. We will analyze these system to collect more samples, and also actually increase the quota value and observe the change of usage distribution in time course.

References

[1] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, 2003.

[2] N. Kobayashi, H. Kuninaka, J. Wakita, and M. Matsushita, "Statistical features of complex systems –toward establishing sociological physics-," *Journal of the Physical Society of Japan*, vol. 80, no. 072001, pp. 1–13, 2011.

[3] H. Kuninaka and M. Matsushita, "Why does zipf's law break down in rank-size distribution of cities?" *Journal of the Physical Society of Japan*, vol. 77, no. 114801, pp. 1–6, 2008.

[4] C. McKnight, "Cost analysis and long term planning over the lefecycle of an enterprise storage solusion," *Journal of Technology Management and Innovation*, vol. 1, no. 5, pp. 87–95, 2006.

[5] E. Ito, Y. Kasahara, and N. Fujimura, "Implementation and operation of the Kyushu University authentication system," in *Proceedings of ACM SIGUCCS'13*, November 2013, pp. 137–142.

[6] N. Fujimura, T. Togawa, Y. Kasahara, and E. Ito, "Introduction and experience with the primary mail service based on their names for students," in *Proceedings of ACM SIGUCCS'12*, November 2012, pp. 11–14.

[7] The R project for statistical computing, http://www.r-project.org/ .

[8] Y. Kasahara, E. Ito, and N. Fujimura, "Introduction of new Kyushu University primary mail service for staff members and students," in *Proceedings of ACM SIGUCCS'14*, no. 103-106, November 2014.