File Size Distribution Model in Enterprise File Server toward Efficient Operational Management

Toshiko Matsumoto, Takashi Onoyama, and Norihisa Komoda

Abstract—**Toward efficient operational management of enterprise file server, we propose an estimation method for relationship between file number and cumulative file size in descending order of file size based on a model for file size distribution. We develop the model by weighted summation of multiple log normal distribution based on AIC. File size data from technical and non-technical divisions of a company show that our model fits well with observed distribution, and that the estimated relationship can be utilized for cost-effective operational management of file server.**

Index Terms— **enterprise file server, file size, operational management, tiered storage**

I. Introduction

Recently enterprise file server has received more and more attention, because of growing trend of unstructured files. In addition, file server is expected to handle archive need for e-Discovery, and real-time backup for Business Continuity Management. From a viewpoint of cost effectiveness, several solutions have been provided: tiered storage technology, de-duplication, and deleting unnecessary files [1], [2]. Tiered storage technology contributes not only to cost-effectiveness of high performance storage system but also to reducing cost for real-time backup in Business Continuity Management by separating active and inactive files. De-duplication reduces the size of data volume in a fileserver and thus is able to shorten backup time. Deleting unnecessary files can improve cost-performance of file servers and leads to more desirable usage of files. Toward theoretical evaluation of algorithms implemented in these solutions, several statistical characteristics of files are reported [3]–[10]. They measured frequencies of file access, extension distribution, and fraction of duplicated contents. Some of them also suggested similarities between file size distribution and Pareto or log normal distribution [4], [8]. However, quantitative estimation requires a survey where

1) system files are discriminated from user files,
2) files are stored in a shared file server for collaboration,
3) data are from industrial file servers, and
4) models are evaluated by statistically testing goodness of fit.

In this study, we propose a model for file size distribution, which is one of the most fundamental statistical characteristics of files. The model is calculated as a weighted summation of multiple log normal distributions. Number of log normal distributions is decided based of Akaike's Information Criterion (AIC) to prevent over fitting [11]. We also describe efficiency estimation for introducing process of tiered storage technology and for reduction of usage volume based on our model. Finally, the model and the estimation are evaluated with actual data.

II. About Data

We use data of shared file servers from 24 divisions of a company: 11 are technical divisions such as research and development, and the other 13 are non-technical divisions such as sales, marketing, and management. Some divisions have tens of thousands of files and others have more than one-million files. Sum of file sizes are between several gigabytes to several hundreds of gigabytes.

Sizes of files in a shared file server of Research and Development division of a company are plotted in Fig 1. X-axis and y-axis represents number and size of a file, respectively. This graph is equivalent to transposed cumulative probability distribution, and demonstrates that there are very few large files. All the other divisions show the same tendency. Fig 2 shows the same data of Fig 1 and data of Pareto distribution [8] in logarithmic scale. Since Pareto distribution is proposed for file size which is larger than some threshold, linear regression line of logarithm value of file size upon logarithm value of file number is calculated with Finney's correction method [12] from plot with file size of 100KB or larger. Fig 3 shows file size histogram of the observed data and of two theoretical distributions proposed in previous studies [4], [8]. Number of classes of the histogram

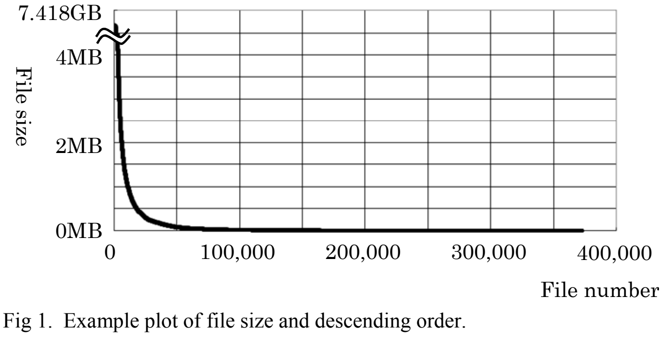

Fig 1. Example plot of file size and descending order.

Manuscript received June 6th, 2012; revised July 2nd, 2012.

T. Matsumoto is with Hitachi Solutions, Ltd., 4-12-7, Higashishinagawa, Shinagawa-ku, Tokyo, 140-0002, Japan (phone: +81-3-5780-2111; fax:+81-3-5780-2049; e-mail: toshiko.matsumoto.jz@hitachi-solutions.com).

T. Onoyama is with Hitachi Solutions, Ltd., 4-12-7, Higashishinagawa, Shinagawa-ku, Tokyo, 140-0002, Japan (e-mail: takashi.onoyama.js@hitachi-solutions.com).

N. Komoda is a professor at Osaka University, 2-1, Yamadaoka, Suita, Osaka 565-0871, Japan (e-mail: komoda@ist.osaka-u.ac.jp).

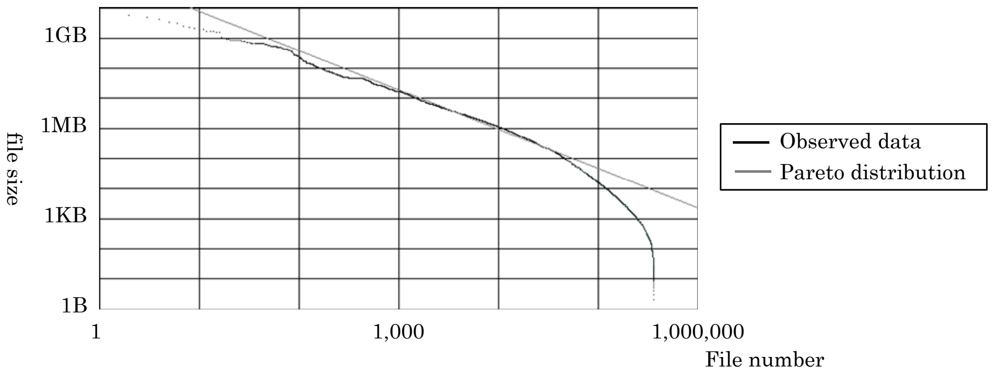

Fig 2. Example plot of file size in descending order for observed data and Pareto distribution.

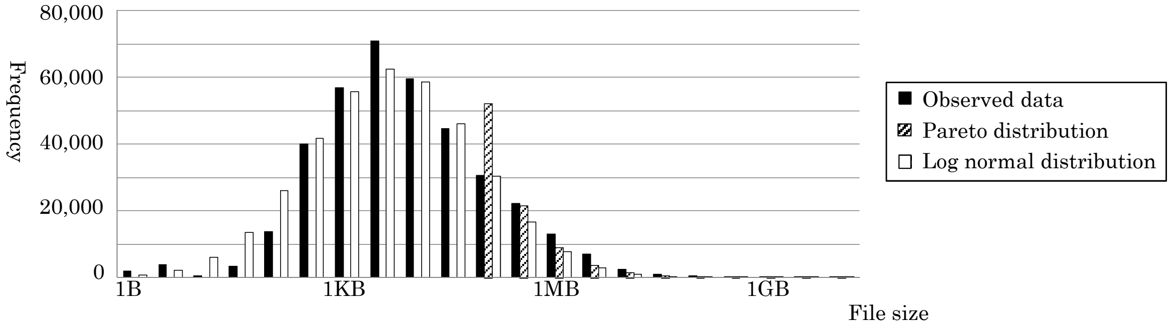

Fig 3. Example plot of file size histogram for observed data and models in previous studies.

is decided based on Sturges' formula [13]. Log normal distribution is calculated by average and standard deviation of logarithm value of file size. Observed plot seems to have linearity in Fig 2 and shows roughly bell-shaped form Fig 3. However, frequency of observed data differs from Pareto distribution by more than 20,000 at 100 KB and from log normal distribution by almost 10,000 at peak in the histogram. Because of these apparent discrepancies, observed histogram shows statistically significant difference both from Pareto distribution and from log normal distribution with type I error rate of 0.01. Even if threshold is set to 1MB, Pareto distribution still shows statistically significant difference from observed data. In all other divisions, observed data differ significantly both from Pareto distribution and from log normal distribution. Therefore, neither Pareto distribution nor log normal distribution fits to observed data, even though they have roughly similar histogram form.

III. File Size Modeling by Weighted Summation of Multiple Log Normal Distributions

A. AIC-based Modeling

We propose file size modeling by weighted summation of multiple log normal distributions. Our model is based on observations where file size distribution depends on content type, such as movie files, database files, and executable files are larger than that of other files [3], [5], and based on an idea that more variety should be observed in general-purpose industrial file servers. Weighted summation of c log normal distributions are calculated to fit to observed data. Similarity is evaluated with χ^2 value from contingency table. χ^2 value is calculated for various c to minimize AIC. AIC is defined as expression (1), where L is maximum likelihood and k is number of degrees of freedom [11].

$$\mathrm{AIC} = 2 \ln L + 2k \tag{1}$$

Likelihood ratio λ is L/L', where L and L' are maximum likelihood values under null hypothesis and in parameter space. AIC can be minimized by minimizing (χ^2 + 3c), because of following three facts. First, L' can be treated as a constant value when observed data is given. Second, λ eventually closes to χ^2 distribution as sample size grows. Third, each log normal distribution adds three degrees of freedom: average, variance, and weight. We can prevent over fitting during model calculation, since AIC value is increased when too many parameters are adopted.

B. Comparison of Observed Data and Proposed Model of File Size

According to the previous section, theoretical distribution of logarithmic scaled file size is calculated as expression (2) where N(μ, σ^2) is normal distribution with average μ and variance σ^2.

$$\begin{aligned}
&281960.8 \times N(3.28,\ 0.60)\\
&+42195.9 \times N(5.01,\ 0.10)\\
&+13121.3 \times N(5.95,\ 0.014)\\
&+12342.1 \times N(2.14,\ 0.028)\\
&+8620.5 \times N(4.23,\ 0.0081)\\
&+7144.9 \times N(6.50,\ 0.23)\\
&+3604.6 \times N(-0.01,\ 0.040)\\
&+2156.7 \times N(0.28,\ 0.0049)\\
&+844.5 \times N(1.38,\ 0.014)\\
&+252.4 \times N(7.88,\ 0.0025)\\
&+141.5 \times N(8.50,\ 0.090)\\
&+75.5 \times N(9.50,\ 9.00)
\end{aligned} \tag{2}$$

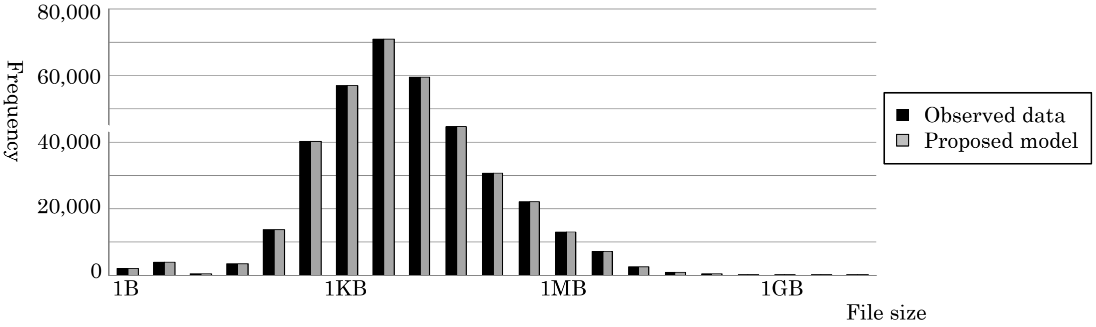

Fig 4. Example plot of file size histogram for observed data and proposed model.

Fig 4 shows file size histogram for observed data and the proposed model. Since theoretical distribution fits very well to the observed data at all classes, no statistically significant difference was observed. Expression (2) shows decrease of weights in an exponentially fashion. The decrease indicates that first term has a dominant effect to the distribution and supports that observed distribution shows roughly bell-shaped form. First term includes source code files and plain text files, that occupy large part of the files in the file server. Second term includes HTML files and PDF files. HTML files have larger size than XML files do in average, because HTML files include specification documents of a programming language and files that are converted by a word processing software.

Observed distributions of file size in all the other 23 divisions fit very well to their corresponding models and demonstrate no statistically significant difference.

IV. Efficiency of Operational Management of File Server Based on Model of File Size Distribution

A. Cases of Operational Management utilizing Model of File Size Distribution

File size distribution model can be directly utilized in the following two kinds of estimation for efficient operational management of file servers. First case is estimating effect of file moving policy in introducing process of tiered storage technology. Processing time of moving file depends both on size summation and on number of files. From these dependences, we can expect that assignment of a high priority to a large file achieves in an efficient moving policy: large volume is moved to lower cost storage in short time and that benefit of tiered storage technology can be realized efficiently with the policy. Since last access time and file size show no correlation (correlation coefficient < 0.1 for all divisions in our data), file size distribution model in the section III A is expected to be effective to estimate relationship between number and size summation of files moved to lower cost storage.

Second case is estimating effect of deleting unnecessary files by users. When millions of files are stored in a file server, it is obviously unrealistic to manually check deletability of all files one by one. Reduction of unnecessary files can be tractable only when large volume is deleted with manual confirmation of a small fraction of files. Deletability confirmation in descending order of file size is effective for efficient volume reduction when the files are deletable. When some files are confirmed to be undeletable, this confirmation policy is still effective for efficient estimation of upper limit of eventual reducible volume.

B. Estimating Efficiency Based on Cumulative File Size

When top n% of large files occupy p% of the total volume, relationship between n% and p% is equivalent to volume efficiency of introducing process of tiered storage technology where inactive files are moved to lower cost storage. The relationship also represents upper limit of reduction volume per confirmation number of file deletability. We can estimate the relationship by integrating theoretical distributions described in section III A.

V. Experimental Result and Evaluation

A. Relationship between number of files and cumulative file size

From expression (2) in section III, Fig 5 shows relationship between fraction of file number n% and fraction of cumulative file size p% in descending order of file size. Value of p% rapidly reaches almost 100%, whereas p% should be equal to n% in randomized order. Rapid increase of p% results from tiny fraction of large files as shown in Fig 1. Since small value of n% can give large p% value, policy with file size descending order is expected to achieve efficient operational management in the both two cases of section IV A.

B. Accuracy Evaluation of Estimating Relationship between Number and Cumulative Size of Files

Fraction of cumulative file size p% is calculated for fraction of file number n% = 1%, 5%, and 10% in 24 divisions. Fig 6, 7, and 8 show comparison results of observed data and estimated value by file size distribution models of Pareto distribution, log normal distribution, and our model of expression (2) in section III. The plots show strong correlation between observed and estimated values of our model in Fig 8 (correlation coefficient ≥ 0.9). In contrast,

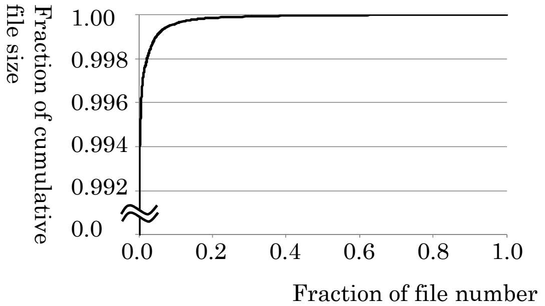

Fig 5. Fraction of file number and of cumulative file size in descending order based on our model.

no apparent correlation is observed for the case of Pareto distribution or log normal distribution (correlation coefficient ≤ 0.6). These results demonstrate that our model can estimate relationship between n% and p% more accurately than models of previous studies can, and is suitable for quantitative evaluation.

C. Shared File Servers in Technical/non-technical Divisions

We compare technical and non-technical divisions regarding fraction of cumulative file size p% for fraction of file number n% = 1%, number of log normal distributions c calculated according to section III A, and overview statistics of file server as shown in Table 1. T-test with Bonferroni correction [14] reveals only p% shows statistically significant difference between technical and non-technical divisions. Values of p% in technical and non-technical divisions are shown in Fig 9. These results mean that overview statistics of file server are not good at estimating p%, although p% depends on type of division. They also indicate that our model can estimate p% better.

VI. Conclusion

In this study, we propose a file size distribution model in enterprise file server and describe that the model is effective for efficient operational management of file servers. Data of shared file servers from various technical and non-technical divisions demonstrate that file size distribution can be modeled as a weighted summation of multiple log normal distributions. The data also demonstrate that integrating the theoretical distribution gives accurate estimate for efficiency of policy with file size descending order in introducing process of tiered storage technology and in deleting unnecessary files. Our estimation can be applied to each file server by two conversions: 1) converting cumulative file size into cost effect on the basis of price and volume capacity of a file server product, and 2) converting file number into introducing process time on the basis of data transfer rate and metadata writing speed in the case of tiered storage technology, and into labor cost on the basis of work efficiency of manual confirmation in the case of deleting unnecessary files. Therefore our model contributes cost-effective file server from viewpoint of operational management. We believe that our model also can be utilized in many other simulation-based evaluations, such as access performance and file fragmentation [10], [15]. Furthermore, file access frequencies, fraction of duplicated content, or more other statistical characteristics can be modeled and be utilized for further efficiency of enterprise file servers.

TABLE I
Overview Statistics of File Server

No.	Overview statistics of file server
1	File number
2	File number (logarithmic value)
3	Sum of file sizes
4	Sum of file sizes (logarithmic value)
5	Average file size
6	Average file size (logarithmic value)
7	Maximum file size
8	Maximum file size (logarithmic value)
9	Maximum file size (logarithmic value) / file number (logarithmic value)

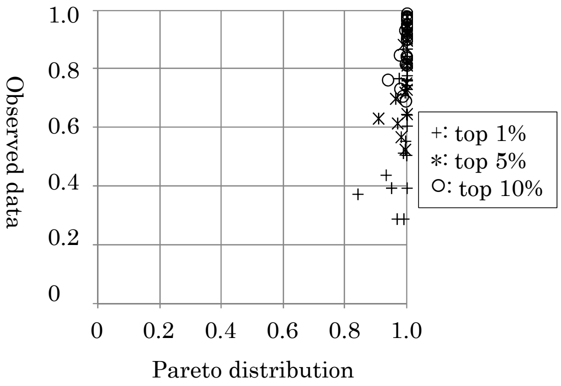

Fig 6. Fraction of cumulative file size occupied by top large files in observed data and in Pareto distribution.

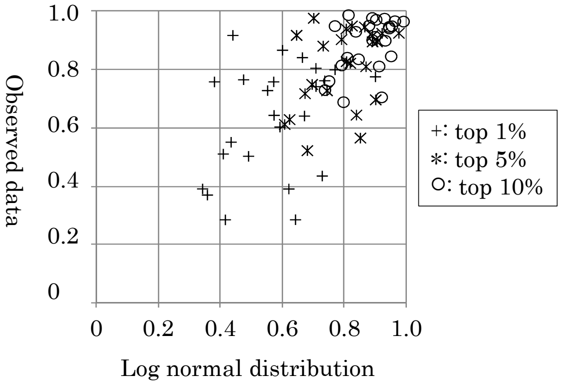

Fig 7. Fraction of cumulative file size occupied by top large files in observed data and in log normal distribution.

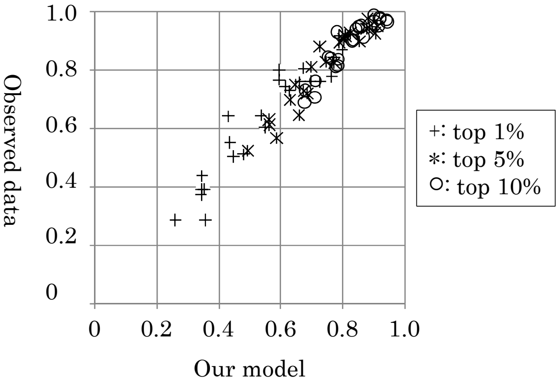

Fig 8. Fraction of cumulative file size occupied by top large files in observed data and in our model.

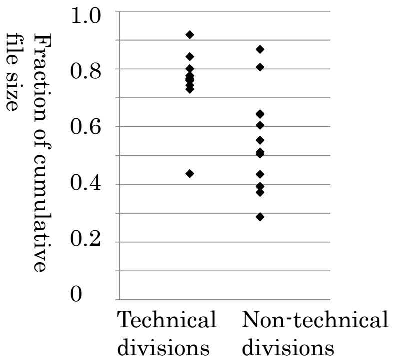

Fig 9. Fraction of cumulative file size occupied by top 1% of large files in technical and non-technical divisions.

References

[1] E. Anderson, J. Hall, J. Hartline, M. Hobbs, A. R. Karlin, J. Saia, R. Swaminathan, and J. Wilkes, "An experimental study of data migration algorithms," in *Proc. of the 5th International Workshop on Algorithm Engineering*, pp.145–158, 2001.

[2] J. Malhotra, P. Sarode, and A. Kamble, "A review of various techniques and approaches of data deduplication," *International Journal of Engineering Practices*, vol. 1, no. 1, pp.1–8, 2012.

[3] N. Agrawal, W. J. Bolosky, J. R. Douceur, and J. R. Lorch, "A five-year study of file-system metadata," *ACM Transactions on Storage*, vol. 3, no.3, pp. 31–45, 2007.

[4] A. B. Downey, "The structural cause of file size distributions," in *Proc. of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, 2001.

[5] K. M. Evans and G. H. Kuenning, "A study of irregularities in file-size distributions," in *Proc. of International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, 2002.

[6] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," in *Proc. of 9th USENIX Conference on File and Storage Technologies*, 2011.

[7] M. Satyanarayanan, "A study of file sizes and functional lifetimes," in *Proc. of the eighth ACM symposium on Operating systems principles*, pp.96–108, 1981.

[8] P. Barford and M. Crovella, "Generating representative web workloads for network and server performance evaluation," in *Proc. of the 1998 ACM SIGMETRICS joint international conference on measurement and modeling of computer systems*, 1998.

[9] T. Gibson and E. L. Miller, "An improved long-term file-usage prediction algorithm," in *Proc. of Annual International Conference on Computer Measurement and Performance*, 1999.

[10] SPEC SFS 2008 benchmark. http://http://www.spec.org/sfs2008/

[11] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. of the 2nd International Symposium on Information Theory*, pp.267–281, 1973.

[12] D. Finney, "On the distribution of avariate whose logarithm is normally distributed," *Journal of Royal Statistical Society*, Supplement VII, pp.155–161, 1941.

[13] H. A. Sturges, "The choice of a class interval," *Journal of American Statistical Association*, vol.21, no.153, pp.65–66, 1926.

[14] G. R. Miller, "Simultaneous statistical inference 2nd edition," New York: Springer-Verlag, 1981.

[15] T. Nakamura and N. Komoda, "Size adjusting pre-allocation methods to improve fragmentation and performance on simultaneous file creation by synchronous write," *ISPJ journal*, vol. 50, no. 11, pp.2690–2698, 2009.