PAPER

Sensitivity Analysis in Optimal Design for Distributed File Allocation Systems

Akiko NAKANIWA[†], **Masaki ONISHI**[†], ***Student Members***, **Hiroyuki EBARA**[†],
and* Hiromi OKADA**[†], ***Regular Members

SUMMARY In distributed network systems, it is one of the most important problems how to assign the files to servers in view of cost and delay. It is obvious that there is a trading-off relationship between costs and delays in these systems. In order to evaluate the optimization that the total cost is minimized subject to the total delay, we have presented the Optimal File Allocation Model as 0–1 integer programming, and have investigated the general characteristics in distributed systems. In this model, we have introduced many cost and delay parameters to evaluate the total cost and delay in the system more exactly. In constructing practical systems, it is necessary to investigate the weight and the contribution of each parameter to the total cost. It is very useful to show how to estimate cost and delay parameters on the basis of this analysis. In this paper, we analyze the sensitivity of these parameters and make clear the influence between principal parameters.

key words: *distributed processing, distributed database system, file allocation, sensitivity analysis, multimedia information network*

1. Introduction

Nowadays, it has been usual for many people to take advantage of various information and communication services. There has been a great trend to shift from centralized systems to distributed systems in view of the reliability, the cost, the delay, the system scalability, and so on. In distributed networks, it is one of the most important problems how to allocate multimedia files in the systems.

There are several servers with multimedia databases in distributed systems. In these distributed systems, files must be stored in at least one of servers, and available for any users. We may easily imagine that it will be preferable to allocate copies of important and popular files in several servers to reduce the communication cost and delay. However, the duplicated storage of the same files in the system have some problems, such as trading-off relationship between the communication cost and the storage cost, the necessity to keep the data consistency, and so on. We should consider the optimal file allocation problem in view of these problems.

C.C. Bisdikian and B.V. Patel presented the optimal file allocation to minimize costs in distributed VoD systems [1], [2]. However, they did not consider delays. It is obvious that there is a trading-off relationship between costs and delays. Also, there are a number of contributions that presented the performance of distributed systems concerning VoD systems [3]–[5]. Our objective is to find the file allocation such that the total cost is minimized subject to the total delay in generic distributed systems.

The authors have presented the optimal file allocation model that may evaluate the optimization from the several viewpoints described above [8]. This model may be applied to various multimedia network environments, such as corporation information networks as well as VoD (Video on Demand) networks, by controlling several parameters. In this model, we have introduced a 0-1 integer programming formulation for the optimization problem. [8], [9] have investigated the general characteristics on file allocation in distributed systems by solving this optimization problem.

In this model, there are many cost and delay parameters affecting the total cost and the total delay, in order to evaluate them in the system more exactly. In constructing practical systems, it is so difficult to investigate how to estimate these cost and delay parameters according to different conditions. Then, it is very useful to examine the weight and the contribution to the total cost of each parameter. That is, it is so meaningful that we examine which parameters have the most significant influence on the system, and what extent the total cost is influenced by the variation of each parameter. We may evaluate the optimal system for various conditions by examining about the behavior of a few parameters which have the most significant influence. So, in this paper, we analyze the sensitivity of these parameters on the Optimal File Allocation Model we have presented, in order to make clear the weight and the contribution of each parameter and estimate values of these parameters.

In this paper, first, we present the optimal file allocation model in Sect. 2, and in Sect. 3, we show the general characteristics on file distribution quantitatively by using an approximate method [9]. In addition, we show the result of the sensitivity analysis in Sect. 4. Finally, we describe our conclusions in Sect. 5.

Manuscript received July 26, 1999.
Manuscript revised November 6, 2000.

[†] The authors are with the Faculty of Engineering, Kansai University, Suita-shi, 564-8680 Japan.

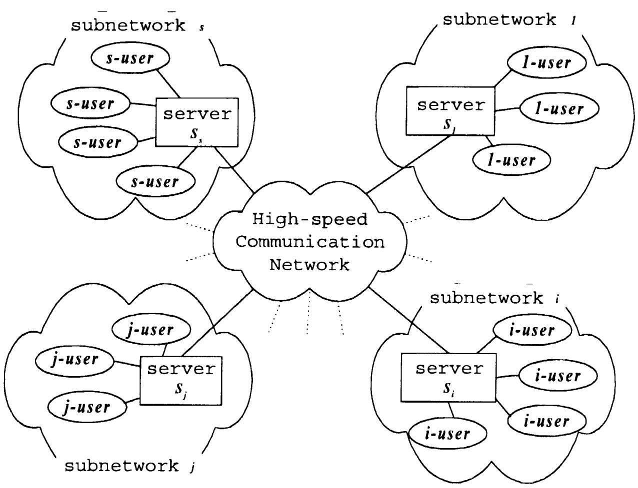


Fig. 1 A model of a distributed information network system.

2. Optimal File Allocation Model [8]

2.1 System Model

We introduce a system model of a distributed information network system that may be applied to a wide range of multimedia network applications. In Fig. 1, we show some typical organization of a distributed information network system. This system consists of servers, users, files, communication lines, and a high-speed communication network.

There exist s servers in the system. Each server is denoted by S_i $(1 \leq i \leq s)$, and has the storage capacity of B_i. Each server constitutes a subnetwork with a number of users. It is assumed that each user of system is connected with one of servers, which is called a local server. The users connected with the server S_i is called *i-users*. In this system, each server is connected by the high-speed communication network, and the whole system works as a distributed database system.

Also, the number of multimedia files is m, and each file is denoted by M_k $(1 \leq k \leq m)$. The maximum number of copies of the file M_k is r_k. It is possible to allocate the same files in several servers, although the number of copies of file M_k is less than or equal to r_k. It is assumed that the size of the file M_k is denoted by F_k. Each server can store these multimedia files so long as the total size of files does not exceed the capacity of the server.

i-users generate file access flow, where the average is λ_i per unit time. This amount of flow generates queuing time at servers. We should assume the queuing process as M/G/1, and here we take M/D/1 for analytical simplicity. The access size from *i-users* for the multimedia files M_k is denoted by L_{ik}. This access size is less than or equal to the file size F_k. Also, let P_{ik} be defined as the access probability from *i-users* for the multimedia files M_k $(\sum_k P_{ik} = 1)$. The parameter P_{ik} determines the traffic matrix in the system. Especially, if the access probabilities have the same value for the same k, the system is file symmetric. Moreover, the probability that *i-users* rewrite the file M_k is denoted by Pw_{ik}. We may discuss various kinds of multimedia network applications according to the value of Pw_{ik}. If we consider VoD system, we must determine that all Pw_{ik} parameters are equal to zero. If Pw_{ik} is large, we may consider that there happens to rewrite the files frequently, such as corporation information networks. In this paper, we apply the write-through consistency protocol to the coherent protocol.

2.2 Formulation

In this paper, we formulate the problem to minimize the total cost per unit time required in the whole system, subject to the total delay in the whole system.

To formulate the optimal file allocation problem, we define 0-1 variables on the allocation of files, X_{ik} [6]. We formulate this optimization problem as 0-1 integer programming by using the following 0-1 variables.

$$X_{ik} = \begin{cases} 1 & (\text{the file } M_k \text{ is in the server } S_i) \\ 0 & (\text{otherwise}) \end{cases} \tag{1}$$

2.2.1 Objective Function

We formulate the total cost per unit time in managing the whole system as the objective function. It is assumed that the total cost C consists of four elements. Those are the communication cost Ct_{ik}, the server cost Crs_{ik}, the storage cost Cst_i, and the line cost Crl_i. The communication cost and the server cost are the costs for each access from an *i-user* for a file M_k. Also, the server S_i is required the storage cost and the line cost per unit time. That is,

$$C = \sum_i \left(\lambda_i \sum_k (Ct_{ik} + Crs_{ik}) + Cst_i + Crl_i \right) \tag{2}$$

We formulate these four elements of the cost on the use of the variables X_{ik}. We may find the optimal file allocation which minimize the total cost, by solving this expression for X_{ik} subject to the restrictive conditions in the next paragraph.

We give full details of the communication cost, the server cost, the storage cost, and the line cost in the following.

[A] Communication Cost Ct
We assume that the communication cost is required to communicate between a user site and his request file site. In this paper, if a file requested from a user is not stored in the local server and several remote servers store the file, the user accesses the server that has the minimum sum of the

communication cost and the server cost. This assumption may cause congestion on a specific server that has the less server cost, if the server cost has more influence than the communication cost. In this case, we may need to assign more copies of files to more servers.

We have the communication cost Ct_{ik} in an access from an *i-user* to a file M_k. Let Ctl_i be defined as the local communication cost coefficient between an *i-user* and a local server S_i. Let Ctr_i be defined as the remote communication cost coefficient between the server S_i and the communication network, which involves the cost to communicate through the communication network.

$$Ct_{ik} = (X_{ik}\mathsf{Ctl}_i + \mathbf{Xf}_{ijk}\mathbf{Ct}_{ij})\, L_{ik}P_{ik} \tag{3}$$

Here, it is assumed that index j equals to the value that minimize the following function on the positive side $(1 \leq j \leq s)$.

$$h1_{min}(j) = \mathbf{Xf}_{ijk}\,(\mathbf{Ct}_{ij} + \mathsf{Csa}_j) \tag{4}$$

Csa_j is the access cost coefficient for a service at the server S_j as described in the next paragraph. Also, we define $\mathbf{Xf}_{ijk}$ and $\mathbf{Ct}_{ij}$ as follows.

$$\mathbf{Xf}_{ijk} = (1 - X_{ik})\, X_{jk} \tag{5}$$

$$\mathbf{Ct}_{ij} = \mathsf{Ctl}_i + \mathsf{Ctr}_i + \mathsf{Ctr}_j \tag{6}$$

[B] Server Cost Crs

We assume that the server cost is the cost of a service for an access from a user for a file at the server. Similarly, if a file requested from a user is not stored in the local server and several remote servers store the file, the user accesses the server that has the minimum sum of the communication cost and the server cost.

We have the server cost Crs_{ik} in an access from an *i-user* to a file M_k. We consider three kinds of costs, that is, the access cost for a service, the rewriting cost to rewrite files, and the relay cost to relay in a local server. The rewriting cost includes the cost of communication in rewriting. Let Csa_i, Csw_i, and Csr_i be defined as the access cost coefficient, the rewriting cost coefficient, and the relay cost coefficient at the server S_i respectively. They depend on the coefficient Crs_i, which is defined as the server cost coefficient of the server S_i, and this coefficient has a fixed value for each server.

$$\begin{aligned} Crs_{ik} &= (X_{ik}\mathsf{Csa}_i + (1 - X_{ik})\mathsf{Csr}_i \\ &\quad + \mathbf{Xf}_{ijk}\mathsf{Csa}_j)L_{ik}P_{ik} \\ &\quad + \left(\sum_{l \neq i}(X_{lk}\mathsf{Csw}_l)\right) RCnf P_{ik}Pw_{ik} \end{aligned} \tag{7}$$

Here, it is assumed that index j equals to the value that minimize the function $h1_{min}(j)$ (Eq. (4)) on the positive side, too $(1 \leq j \leq s)$. Also, $RCnf$ will be defined as the local-global difference of the communication cost in the next section.

[C] Storage Cost Cst

We assume that the storage cost is required for each server to store multimedia files. It is charged on unit time basis, and not dependent on the frequency of accesses. In this paper, the storage cost depends on the size of all files stored at a server. We have the storage cost Cst_i of the server S_i. Let Cst be the storage cost coefficient, which has a fixed value in common for all servers.

$$Cst_i = \mathsf{Cst}\sum_k X_{ik}F_k \tag{8}$$

[D] Line Cost Crl

We assume that the line cost is the cost to use a communication line between each server and the high-speed communication network. This is charged on unit time basis and does not depend on the frequency of accesses, like the storage cost. In this paper, we consider that the line cost Crl_i between the server S_i and the communication network has a fixed value for each server.

2.2.2 Restrictive Conditions

We formulate the total delay in working the whole system as the restrictive conditions for the optimal file allocation problem. It is assumed that the total delay consists of the communication delay Dt_{ik}, the service delay Dx_i, and the rewriting delay Dw_i. For each access from an *i-user* for a file M_k, an *i-user* has to spend these delays. Let D_{max} be defined as the maximum total delay, and the total delay is as follows.

$$\begin{aligned} D &= \sum_i \left(\lambda_i \left(\sum_k Dt_{ik} + Dw_i\right) + \lambda_i^* Dx_i\right) \\ &\leq D_{max} \end{aligned} \tag{9}$$

The total delay is the sum of the communication delay, the service delay, and the rewriting delay for each access from an *i-user* multiplied by the average access times of an *i-user*. Here, for the service delay, we do not use the flow λ_i but λ_i^* as the mean arrival rate. When we consider service in each server, we should consider not only λ_i accesses on average from local users but also accesses from users connected with other servers. Thus, we have λ_i^*.

$$\lambda_i^* = \lambda_i + \sum_{j'} \delta\,(j - i) \sum_k (\mathbf{Xf}_{j'ik}\lambda_{j'}P_{j'k}) \tag{10}$$

Here, it is assumed that j equals to the value that minimize the following function $h2_{min}(j)$ on the positive side $(1 \leq j \leq s)$.

$$h2_{min}(j) = \mathbf{Xf}_{j'jk}\left(\mathbf{Ct}_{j'j} + \mathsf{Csa}_j\right) \tag{11}$$

$$\delta(t) = \begin{cases} 1 & (t = 0) \\ 0 & (t \neq 0) \end{cases} \tag{12}$$

[A] Communication Delay Dt

The communication delay is spent to communicate from a user to a file per an access. Let Dtl_i be defined as the local communication delay coefficient between an *i-user* and a local server S_i. Let Dtr_i be defined as the remote communication delay coefficient between a server S_i and the communication network. We have the communication delay Dt_{ik} in an access from an *i-user* for a file M_k as follows. Here, it is assumed that index j equals to the value that minimize the function $h1_{min}(j)$ (Eq. (4)) on the positive side, too $(1 \leq j \leq s)$.

$$Dt_{ik} = \left(X_{ik}\mathsf{Dtl}_i + \mathbf{Xf}_{ijk}\mathbf{Dt}_{ij}\right) L_{ik} P_{ik} \tag{13}$$

Also, we define $\mathbf{Dt}_{ij}$ as follows.

$$\mathbf{Dt}_{ij} = \mathsf{Dtl}_i + \mathsf{Dtr}_i + \mathsf{Dtr}_j \tag{14}$$

[B] Service Delay Dx

The service delay is the delay of a service to a file in a server. We may assume this delay behavior as M/G/1, and here we take M/D/1 for analytical simplicity. We can introduce the service delay Dx_i in a server S_i by the *Pollaczek-Khinchin mean-value formula* and *Little's result* [7]. A service in a server S_i is assumed to take the service time x_i on average.

$$Dx_i = x_i + \frac{\lambda_i^* x_i^2}{2\left(1 - \lambda_i^* x_i\right)} \tag{15}$$

Here, we do not use the flow λ_i but λ_i^* as the mean arrival rate.

[C] Rewriting Delay Dw

The rewriting delay is spent to keep the data consistency, when the same file is allocated in more than or equal to two servers. This rewriting delay is evaluated per a rewriting from a user to a file, and includes a delay of communication in rewriting. In this paper, we apply the write-through consistency protocol in which all of the same files in a system should be modified simultaneously, when one of them is rewritten. The rewriting delay coefficient is denoted by Dw. We have the rewriting delay Dw_i in rewriting from an *i-user* to a file M_k.

$$Dw_i = \sum_k \mathsf{Dw} RDnf\left(W_k - 1 + \delta\left(W_k\right)\right) \cdot P_{ik} Pw_{ik} \tag{16}$$

Here, $RDnf$ will be defined as the local-global difference of the communication delay in the next section. Also, W_k is the number of files in the whole system, and denoted as follows.

$$W_k = \sum_i X_{ik} \tag{17}$$

3. General Characteristics

In Sects. 3 and 4, we show numerical results of the optimal file allocation problem presented in the previous desection. First, we show the general characteristics in distributed systems quantitatively by the numerical results in this section. In the next section, we analyze the sensitivity of the cost and delay parameters.

3.1 Parameters for Numerical Results and Considerations

Considering the optimal file allocation problem, we define the distribution rate Rd for performance measure as follows.

$$Rd = \frac{\sum_i \sum_k X_{ik}}{m} \tag{18}$$

Rd is the ratio of the number of all files allocated in the system to the number of kinds of files.

We obtain the assignment of files to servers as a solution. On the basis of this assignment, we calculate distribution rates, the total cost, the total delay, and so on. We describe our considerations about optimization results according to these criterion.

Moreover, we define the other parameters that deal with the economic environments of the distributed systems as follows.

- $RCnf$: The local-global difference of the communication cost. This parameter means the difference between the local communication cost from a user to its local server and the global communication cost from a local server to a remote server stored a request file. The larger a local-global difference $RCnf$ is, the more increasing a difference of the communication cost between to a local server and to remote servers is.
- $RDnf$: The local-global difference of the communication delay. This parameter means the difference between the local communication delay from a user to its local server and the global communication delay from a local server to a remote server stored a request file.

3.2 General Characteristics in Distributed Systems

Here, we show the optimization results for the above problem. In this paper, we use the approximate method for the optimization problem to deal with large-scale systems. The approximate algorithm we use is based on the Greedy Method. The Greedy Method finds feasible solutions directly fixing each variable in accordance

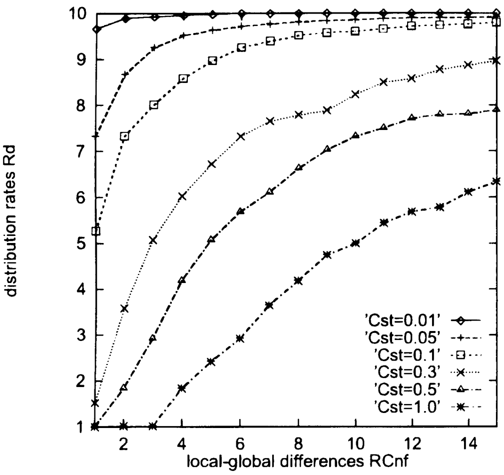


Fig. 2 A trading-off between the storage costs and the communication costs ($RDnf = 5.0, Pw = 0.1$).

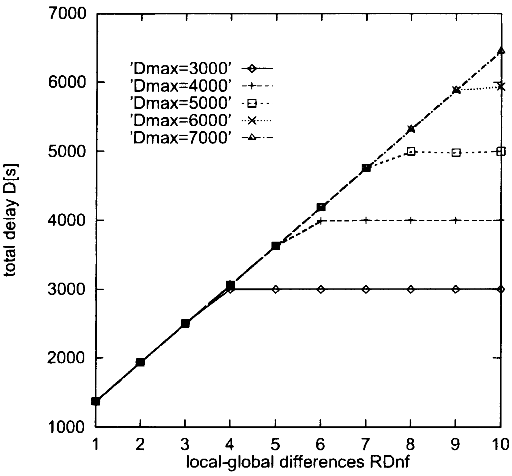


Fig. 3 A delay characteristics vs. $RDnf$ ($RCnf = 3.0, Pw = 0.1, \mathsf{Cst} = 0.5$).

with a local critical value for contribution to an objective function. In our approximate method, we take the access frequency to the files as the local critical value, that is $\lambda_i \times P_{ik}$. We present the approximate method, and make a comparison between this approximate method and the exhaustive search in [9].

We show the general characteristics in distributed systems quantitatively by using this approximate method. Here, we consider the system parameters as follows. There exist 10 servers ($s = 10$), and the capacity of each server is 100 [Gbyte] ($B_i = 100000, 1 \leq i \leq 10$). Also, there exist 1000 kinds of multimedia files ($m = 1000$), and the size of each file is 100 [Mbyte] ($F_k = 100, 1 \leq k \leq 1000$). The maximum number of copies of a file M_k is less than 10 ($r_k = 10$).

In case we use the exhaustive search, system scales that we can deal with is limited to the system which has 4 servers and 6 kind of files. Though the above system scale does not cover all actual systems, we consider it is enough scale which cover many systems. In compensation for increasing in computing time, we can deal with more large scale of system. Also, in this paper, we consider system parameters, such as the capacity of servers, the size of files, and so on, are identical for analytical simplicity. However, we may deal with many kinds of actual multimedia networks by controlling these system parameters in our model.

In Fig. 2, we show the trading-off relationship between the storage costs and the communication costs. This figure shows the distribution characteristics vs. the local-global differences of the communication cost $RCnf$ for different values of the storage cost coefficient Cst, i.e., 0.01, 0.05, 0.1, 0.3, 0.5, 1.0. As shown in Fig. 2, when the storage cost coefficient Cst is larger, the distribution rates Rd become smaller. When the storage costs are large, files should be centralized. The smaller the storage costs is, the more files tend to be distributed to minimize the total cost. On the other hand, as shown in Fig. 2, when $RCnf$ becomes larger, more files tend to be distributed. As $RCnf$ grows larger, the communication costs to remote server become increasing. Thus, files become to be distributed and the distribution rates become increasing, so that files will be declined necessity to communicate and the communication costs will be reduced.

Next, we show the effect of the delay restrictions on the optimization of file allocation. In Fig. 3, we show the characteristics of the total delay vs. the local-global differences of the communication delay $RDnf$ for different values of the maximum total delay D_{max}, i.e., 3000, 4000, 5000, 6000, 7000. We see the total delay are not affected by the restriction when $D_{max} = 7000$. However, when $D_{max} = 3000, 4000, 5000, 6000$, the restriction becomes to have influence on the total delay in case $RDnf$ becomes larger. In this case, the total delay may exceed the maximum in the optimal file allocation that minimize the total cost. We show this effect of the delay restriction on the file allocation in Fig. 4. In case that $RDnf$ is larger value, files are more distributed to reduce the communication delays. This figure shows that files should be distributed, when the maximum total delay is small.

4. Sensitivity Analysis

In the model presented here, there are many parameters affecting the total cost and the total delay, in order to deal with the system more exactly. When we construct practical systems, it is very important to show how to estimate all these parameters. We need to examine the weight and the contribution to the total cost of each parameter. That is, we should investigate which parameters have the most significant influence, and what extent the total cost is influenced by the variation of each parameter. In this section, we analyze the sensitivity of these parameters on file allocation in distributed

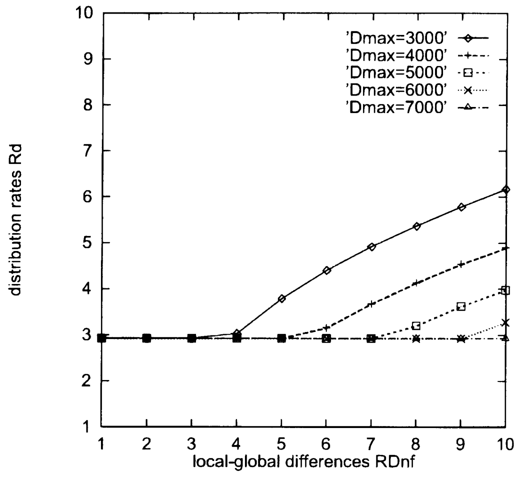


Fig. 4 A distribution characteristics vs. $RDnf$ ($RCnf = 3.0, Pw = 0.1, \mathrm{Cst} = 0.5$).

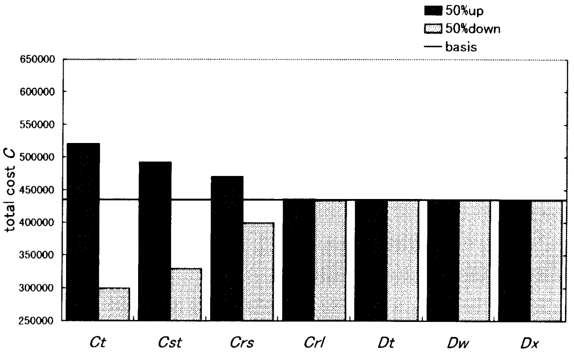


Fig. 5 Sensitivity of cost and delay parameters ($D_{max} = 8000$).

networks.

4.1 Sensitivity of Cost and Delay Parameters

First, we show the comparison of sensitivity of the cost and delay parameters defined above. Figure 5 shows the total cost for the case of both 50% increasing and 50% decreasing to the basic value in each parameter to investigate its influence. In this figure, the line of basis is the total cost in case that all cost and delay parameters have the basic values. The sensitivity of each parameter is given by the difference between the basis and the case of the variation of 50% of each parameter. As shown in this figure, the variations in the communication cost Ct and the storage cost Cst have significant influence on the objective function. Also, the delay parameters hardly have influence on the file allocation. We can easily imagine there is hardly effect of the delay restriction, since we set the maximum total delay $D_{max} = 8000$ in this figure.

We show the same comparison results in case the delay restrictions have influence in Fig. 6 and Fig. 7. In these figures, we set $D_{max} = 4000, 2000$ respectively.

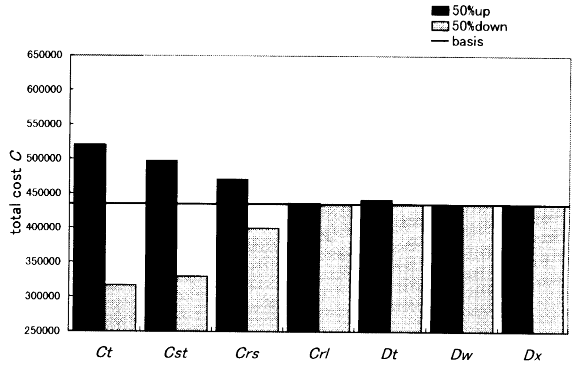


Fig. 6 Sensitivity of cost and delay parameters ($D_{max} = 4000$).

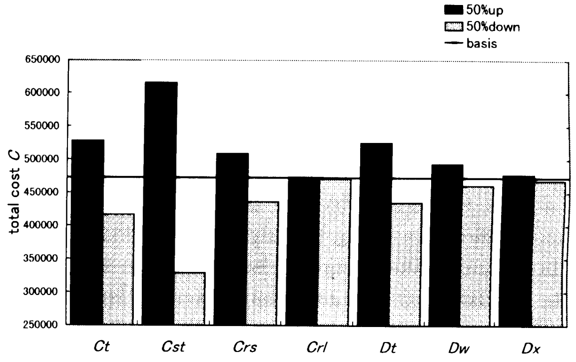


Fig. 7 Sensitivity of cost and delay parameters ($D_{max} = 2000$).

As shown in Fig. 6, the communication delay Dt becomes to have a little influence on the objective function in case of 50% increasing, in case of $D_{max} = 4000$. Moreover, in case of $D_{max} = 2000$, where the restrictive condition is more tight, the communication delay Dt becomes to have much more influence. Besides, the rewriting delay Dw and the service delay Dx become to have influence. Also, in Fig. 7, the line of basis itself becomes to have a larger value than the case of Fig. 6 and Fig. 7. These results show the trading-off relationship that the total cost is increasing when the delay is restrained lower. We may estimate that the influences of delay parameters do not become sensitive furthermore according as the delay restriction becomes tight, since there is the tendency that the line of basis also becomes upper.

There is another remarkable results in these figures. As the delay restriction becomes tight, the storage cost Cst becomes to have much more influence on the total cost than Ct. We give full detail of results for the change of the delay restriction in Sect. 4.3.

As shown in these results, the sensitivity of cost parameters is more than delay parameters. In the next section, we concentrate on the sensitivity of cost parameters.

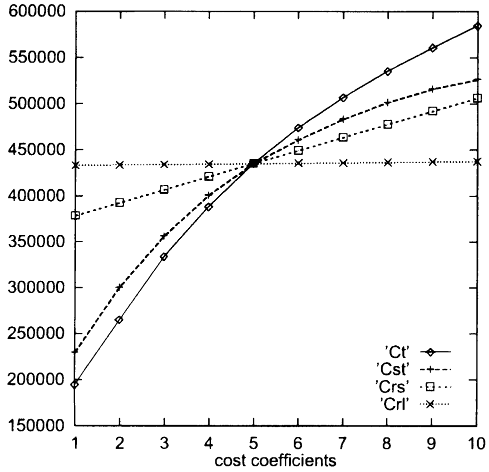


Fig. 8 A total cost characteristics in case of $D_{max} = 8000$ vs. cost parameters, the local communication cost coefficient Ctl times 5, the storage cost coefficient Cst times 10, the server cost coefficient Crs times 5, the line cost coefficient Cl times 1/50 ($RCnf = 3.0, RDnf = 5.0, Pw = 0.1$).

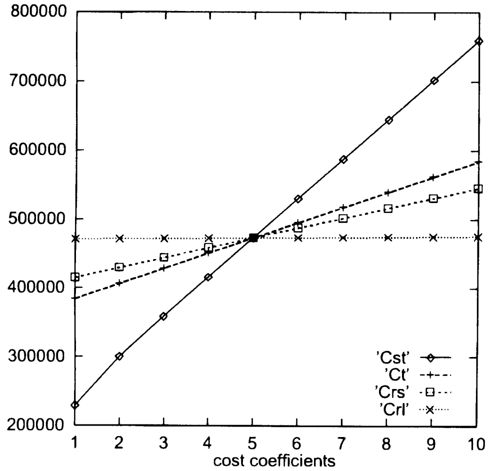


Fig. 9 A total cost characteristics in case of $D_{max} = 2000$ vs. cost parameters, the storage cost coefficient Cst times 10, the local communication cost coefficient Ctl times 5, the server cost coefficient Crs times 5, the line cost coefficient Cl times 1/50 ($RCnf = 3.0, RDnf = 5.0, Pw = 0.1$).

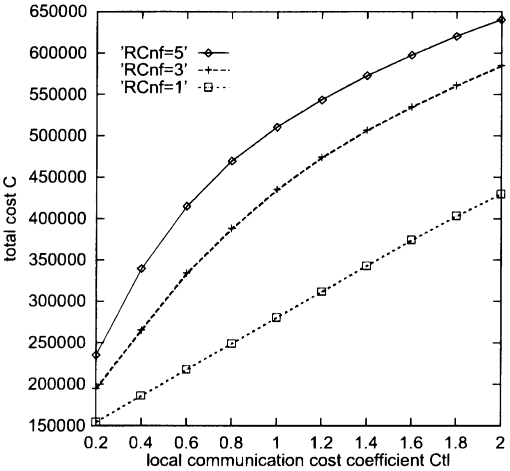


Fig. 10 A total cost characteristics vs. local communication cost coefficient Ctl ($D_{max} = 8000, RDnf = 5.0, Pw = 0.1$).

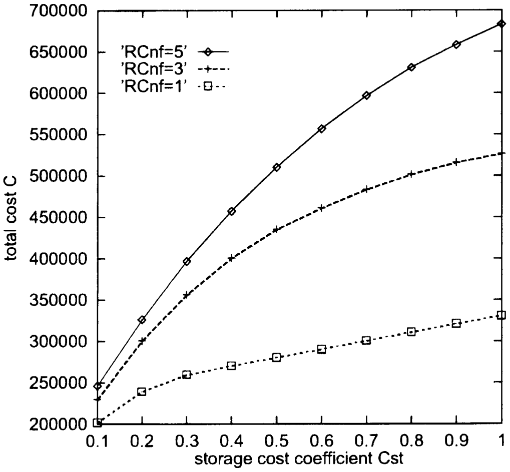


Fig. 11 A total cost characteristics vs. storage cost coefficient Cst ($D_{max} = 8000, RDnf = 5.0, Pw = 0.1$).

4.2 Sensitivity of Cost Parameters

We show the sensitivity of the cost parameters in detail. Figure 8 and Fig. 9 show the total cost characteristics vs. the variation of the cost parameters for $D_{max} = 8000$ and $D_{max} = 2000$ respectively. These figures is normalized at the point the cost coefficients is equal to 5. When the cost coefficients is equal to 5, all lines cross one another in these figures. These points are corresponding to the line of basis in Fig. 5 and Fig. 7. The gradient of each line in these figures means the sensitivity of each cost parameter. As shown in Fig. 8, the order of sensitivity is Ct, Cst, Crs, Crl, in case of $D_{max} = 8000$. And in case of $D_{max} = 2000$, the order is Cst, Ct, Crs, Crl. That is, we also see in these figures that the storage cost Cst becomes to have more influence on the total cost than Ct as the delay restriction becomes tight (see Sect. 4.3). In addition, as we may see in these figures, the total cost grows almost linearly according to the increase in all cost parameters. We may estimate the value of these cost parameters according to this tendency.

From both results, we can see the communication cost and the storage cost is the most sensitive parameters. Thus, we show the sensitivity of communication cost and the storage cost in detail.

In Fig. 10 and Fig. 11, we show the total cost characteristics for different values of the local-global differences of the communication cost $RCnf$, i.e., 1, 3, 5, vs. the communication cost and the storage cost respectively. As we may also see in these figures, the sensitivity of the communication cost and the storage cost have linear characteristics. This tendency becomes more clear for large values of both cost parameters. Also, in Fig. 10, the gradients of lines is almost identical for different values of the local-global differences *RCnf*. On the other hand, the gradient of the line becomes large when the local-global differences *RCnf* is larger in Fig. 11. As *RCnf* grows larger, the files should be distributed at many servers and the storage cost becomes to have more significant influence, that is, the storage cost becomes more sensitive.

4.3 Influence of Delay Restrictions

Finally, we give full detail of the influence of the delay restrictions on the sensitivity of each paramter, which is appeared in the above analysis. In Fig. 12 and Fig. 13, we show the sensitivity of the communication cost Ct and the storage cost Cst for the change of the delay restriction. These figures show the total cost characteristics for different values of the maximum total delay D_{max} (2000, 4000, 8000), vs. the communication cost and the storage cost respectively. As shown in Fig. 12, the gradient of the line becomes smaller as the maximum total delay D_{max} is smaller, that is, the delay restriction becomes more tight. On the other hand, in Fig. 13, the gradient of line becomes larger as D_{max} is smaller. Files are more distributed to reduce the communication delays when the restriction is tight as described in Fig. 4. Thus, when D_{max} is small, the storage cost becomes to have much more significant influence, while the communication cost becomes to have less influence. Moreover, as we can see in these figures, when the local communication cost coefficient Ctl is more than 1.6, and the storage cost coefficient Cst is less than 0.3, the gradients of three lines are identical. When the communication cost is large and the storage cost is small, files are more distributed. We can imagine that the total delay does not exceed the maximum D_{max} due to this distribution of files in any cases. Thus, there are hardly differences in the gradients for different D_{max}, when the communication cost is large and the storage cost is small.

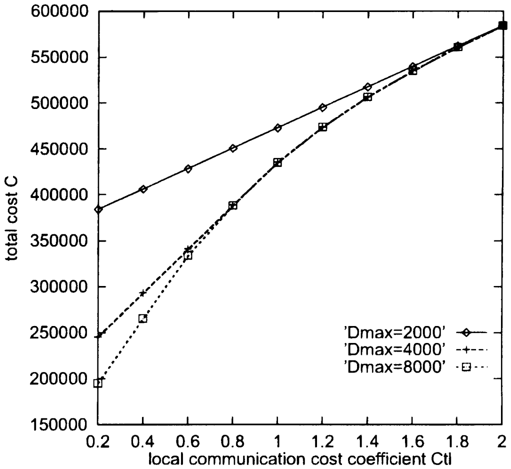


Fig. 12 A total cost characteristics vs. the communication cost for different values of D_{max} ($RCnf = 3, RDnf = 5.0, Pw = 0.1$).

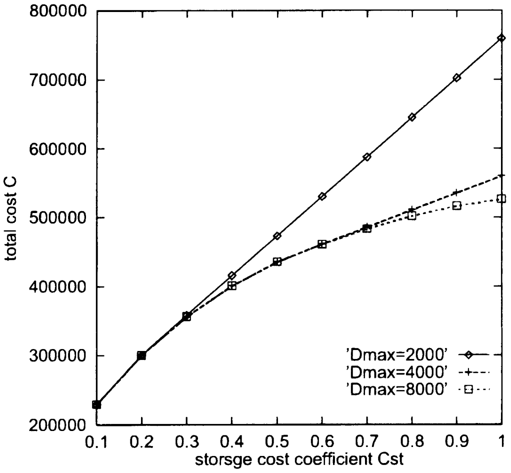


Fig. 13 A total cost characteristics vs. the storage cost for different values of D_{max} ($RCnf = 3, RDnf = 5.0, Pw = 0.1$).

5. Conclusions

In this paper, we have optimized the file allocation, such that the total cost is minimized subject to the total delay in distributed information network system. For considering this optimization problem, we have presented the Optimal File Allocation Model that covers a wide range of multimedia network applications. In this model, we have formulated this problem as a 0-1 integer programming, and we have described the considerations of numerical results by solving this optimization problem.

As the numerical results by the approximate method, we have shown the general characteristics on file distribution in practical large scale of systems, such as a trading-off between the communication costs and the storage costs and the effect of the delay restriction. As shown in numerical results, we may quantify the general characteristics in distributed systems. The system model presented here is very useful for evaluating various elements in constructing a distributed system.

In our model, there are many parameters affecting the total cost and the total delay, in order to deal with the system more exactly. It is very useful to examine the weight and the contribution of each parameter in constructing practical systems. We have analyzed the sensitivity of the cost and delay parameters. As a result, the variation in the communication cost and the storage cost have the most significant influence on the objective function mainly. Also, the sensitivity of these cost parameters have the linear characteristics. According to this characteristics, we can estimate the values of these parameters when we construct practical systems. In addition, we have shown the influence of the delay restrictions on the sensitivity of these parameters. We have made clear that we can grasp the influence among principal parameters in full detail by this sensitivity analysis.

References

[1] C.C. Bisdikian and B.V. Patel, "Issues on movie allocation in distributed video-on-demand systems," IEEE Proc. ICC'95, pp.250–255, 1995.

[2] C.C. Bisdikian and B.V. Patel, "Cost-based program allocation for distributed multimedia-on-demand systems," IEEE Multimedia, pp.62–72, Fall 1996.

[3] D. Deloddere, W. Verbiest, and H. Verhille, "Interactive video on demand," IEEE Commun. Mag., pp.82–88, May 1994.

[4] V.O.K. Li, W. Liao, X. Qiu, and E.W.M. Wong, "Performance model of interactive video-on-demand systems," IEEE J. Sel. Areas Commun., vol.14, no.6, Aug. 1996.

[5] S.A. Barnett and G.J. Anido, "A cost comparison of distributed and centralized approaches to video-on-demand," IEEE J. Sel. Areas Commun., vol.14, no.6, Aug. 1996.

[6] W.W. Chu, "Optimal file allocation in a multiple computer system," IEEE Trans. Comput., vol.C-18, no.10, pp.885–890, Oct. 1969.

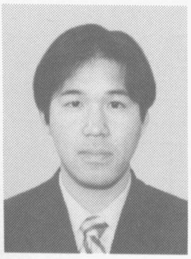
[7] L. Kleinrock, QUEUEING SYSTEMS, pp.180–191, A Wiley-Interscience Publication, 1975.

[8] A. Nakaniwa, H. Ebara, and H. Okada, "File allocation designs for distributed multimedia information networks," IEICE Trans. Commun., vol.E81-B, no.8, Aug. 1998.

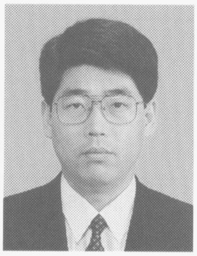
[9] A. Nakaniwa, M. Onishi, H. Ebara, and H. Okada, "File allocation in distributed multimedia information networks," Proc. IEEE GLOBECOM'98, Nov. 1998.

[10] A. Nakaniwa, M. Onishi, H. Ebara, and H. Okada, "Sensitivity analysis of file allocation for distributed information networks," Proc. IEEE ICC'99, June 1999.

Akiko Nakaniwa was born in 1975. She received B.S. and M.S. degrees from Kansai University, Japan, in 1997 and 1999. Currently she is studying in the graduate school of Kansai University. Her research interest is optimization design of distributed database systems. She is a student member of IEEE.



Masaki Onishi was born in 1976. He received B.S. and M.S. degrees from Kansai University, Japan, in 1998 and 2000. He joined NTT WEST in 2000. His research interest is combinatorial optimization. He is a student member of IEEE.



Hiroyuki Ebara was born in 1958. He received the B.S., M.S., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1982, 1984, and 1987, respectively. In 1987 he became Assistant Professor of Osaka University. Since 1994 he has been with Kansai University, where he is currently Associate Professor. His main research interests are in computational geometry, combinatorial optimization, and parallel computing. Dr. Ebara is a member of IEEE, ACM, and SIAM.

Hiromi Okada received the B.S., M.S. and Ph.D. degrees on Communications Engineering, all from Osaka University, Japan, in 1970, 1972, and 1975, respectively. From 1975 to 1983, he was an assistant professor of Osaka university. From 1983 to 1987, he was an associate professor of Kobe university. From 1987 to 1996, he was with Osaka University as an associate professor. Since April 1996, he has been with the department of Electronics Engineering, Faculty of Engineering, Kansai University as a professor. His current research interest includes ATM multicasting networks, wireless ATM networks, ad-hoc wireless networks, performance evaluation and optimization of distributed information networks. He is the author of several books entitled "Information Networks" (in Japanese, Baifukan Pub.), "Introduction of Computer Systems" (in Japanese, Shokodo Pub.) and so on. He is a member of IEEE, IPS (Information Processing Society) Japan.