An Efficient Data Transfer Method for Distributed Storage System over Satellite Networks

---

Citation:

Katsuya Suto, Panu Avakul, Hiroki Nishiyama, and Nei Kato, "An Efficient Data Transfer Method for Distributed Storage System over Satellite Networks," IEEE 77th Vehicular Technology Conference (VTC2013 Spring), Dresden, Germany, Jun. 2013.

URL:

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6691869

An Efficient Data Transfer Method for Distributed Storage System over Satellite Networks

Katsuya Suto, Panu Avakul, Hiroki Nishiyama, Nei Kato
Graduate School of Information Sciences, Tohoku University, Sendai, Japan
E-mails: {suto, panu0014, bigtree, kato}@it.ecei.tohoku.ac.jp

***Abstract*—We study a novel distributed storage system integrated Data Centers (DCs) and satellite networks. This integrated system is expected as distributed storage system that can keep the storage service even if disasters strike because satellite is tolerant to link disruption caused by disasters. In this paper, we focus on data distribution method in the integrated system, and assume an erasure coding and a simply replication as data distribution method. We evaluate the storage volume and transmission time on each method which are required to restore lost data when some DCs are damaged by disasters. The storage volume of the erasure coding becomes lower than that of the replication while the transmission time becomes higher. A data transfer method is proposed in this paper to shorten the transmission time of the erasure coding. The proposed method to reduce the transmission volume on downlink communication by using network coding technologies. The numerical results show that the proposed method can restore the lost data in less time.**

I. Introduction

Nowadays, storage requirements increase almost exponentially due to widespread use of email, photos, videos, log files, and so forth. The distributed storage system have attracted much attention as a fault tolerant storage, because they distributes and stores data in many Data Centers (DCs) [1], [2]. However, the earthquake and tsunami dramatically affected communications infrastructures and made it difficult to provide reliable storage services. The great East-Japan Catastrophic Disaster in March 2011 damaged 1.9 million circuits out of total 24 million and $29,000$ base stations out of total $132,000$ [3]. Under such circumstances, it is possible to identify two problems, which are communication to disaster area and communication (and storage) in disaster area.

To cope with the first problem, SKY Perfect JSAT corporation has developed a novel distributed storage system called S*Plex3 [4], which integrates DCs and satellite networks. The integrated storage system can communicate with disaster area during disasters since satellite networks have large coverage area and high disaster tolerance [5]. Additionally, it is possible to minimize the risk of losing large number of DCs to a single disaster event by distributing them many different physical locations (throughout Japan). Moreover, the integrated storage system can ensure that we can restore the data when we have less than 30% data loss by using an erasure coding as data distribution method, which is basic technology in satellite networks [6]. Therefore, the integrated system can provide the storage service even if some DCs are damaged by disasters. However, disaster does not only damage DCs and communications line to the disaster area but also communications infrastructures in the disaster area such as optical cable, base station, and so forth. Thus, communications and storage in disaster area is also the problem. To cope with the second problem, Nippon Telegraph and Telephone (NTT) corporation is developing a Movable and Deployable Resource Unit (MDRU), which is a vehicle equipped with communications and storage equipments. It is expected as an emergent communications and storage device instead of malfunctioning DCs and base station [7], [8].

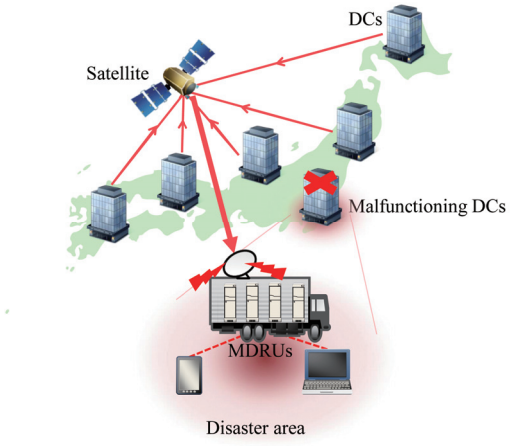


Fig. 1. Considered system model to expeditiously restore the lost data.

We strive for a distributed storage system, which can expeditiously restore the lost data by using the integrated storage system and MDRUs as shown in Fig. 1. The integrated storage system is used to distribute the data to different DCs. When some DCs are damaged by disasters, MDRUs are deployed to the disaster area as a communications and storage infrastructure. Subsequently, remaining DCs transmit the redundant data to MDRUs via satellite networks to restore the lost data. In this paper, we investigate the impact of data distribution method on the network performance such as the storage volume and the transmission time which are required to restore the data when some DCs are damaged by disasters. An erasure coding and a simply replication are evaluate as data distribution method. The storage volume of the erasure coding becomes lower than that of the replication while the transmission time becomes higher. Therefore, we propose a novel method to shorten the transmission time required to restore the lost data in the integrated storage system with

TABLE I
PARAMETERS IN THE SYSTEM MODEL.

Parameter	Definition
$N_{\rm dc}$	Number of DCs
c	Each Link capacity between DCs and satellite
$\overline{s}$	File size
$N_{\rm file}$	Number of files $N_{\rm file} = mN_{\rm dc}, m \in \mathcal{N}^+$
p	Actual number of damaged DCs $D > p \geq 1$
f	Maximum number of damaged DCs $D > f \geq p$

erasure coding. The proposed method reduce the transmission volume on downlink communication by using network coding technologies [9].

The remainder of this paper is organized as follows. Section II introduces the system models, the replication, and the erasure coding as data distribution method considered in this paper, and also analyze and investigate the storage volume and transmission time on each method. We propose the method to shorten the transmission time of the erasure coding in Section III. Section IV provides numerical results and Section V concludes the paper.

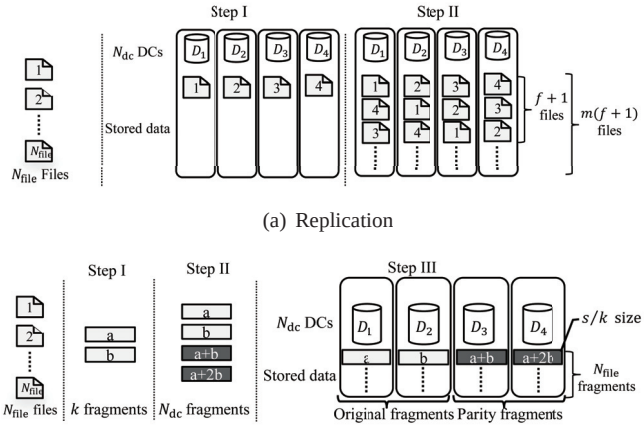
II. SYSTEM MODELS

A. Preliminaries

The integrated storage system consists of a Geostationary Earth Orbit satellite and $N_{\rm dc}$ DCs. We assume that the satellite has a coverage area which can communicate with all of DCs, and employs Frequency Division Multiple Access (FDMA) and Pre-Assigned Multiple Access (PAMA) as bandwidth allocation methods. Thus, the capacity of each link between DCs and the satellite becomes the same value c. Additionally, we assume that uplink and downlink capacity over satellite communications are the same value c, where packet loss (or bit error) is ignored. The integrated storage system distributes $N_{\rm file}$ files with s byte of data to DCs. Supposed that each DC stores same amount of data, $N_{\rm file}$ becomes equal to $mN_{\rm dc}, m \in \mathcal{N}^+$. When p DCs are damaged by disasters, p MDRUs are deploy to disaster area and provide the storage and communications service as a substitute of malfunctioning DCs and base stations. f represents the maximum number of malfunctioning DCs that we can have in order to be able to restore the lost data. In order to not consider the case that files cannot be restored, $f \geq p$ shall be required. We study the performance of two data distribution methods, i.e., the replication and the erasure coding, when the values of f and p are changed. Table I summarizes the parameters and their definition.

B. Replication

Fig 2(a) shows the data distribution procedure to restore lost data when f DCs are damaged in the integrated storage system with replication. First, $N_{\rm dc}$ files are distributed into $N_{\rm dc}$ DCs. Then, f replications on each file are distributed to f DCs, here each DC has same number of files. Such procedure continues until all of files are distributed, namely m times. Therefore, the storage volume on each DC, $S_{\rm rep}$, is defined by the following



(a) Replication

(b) Erasure coding

Fig. 2. Example of data distribution Procedures when $f = 2$.

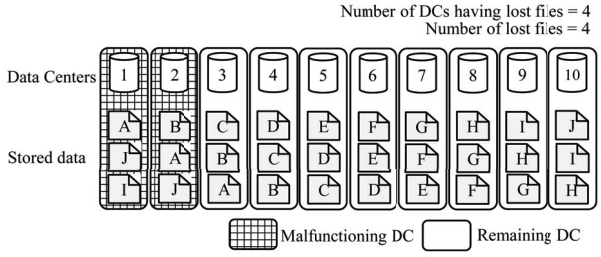
equation, because it is necessary for each DC to store $m(f+1)$ files as shown in Fig 2(a).

$$S_{\rm rep} \quad = \quad \overline{s}m(f+1), \tag{1}$$

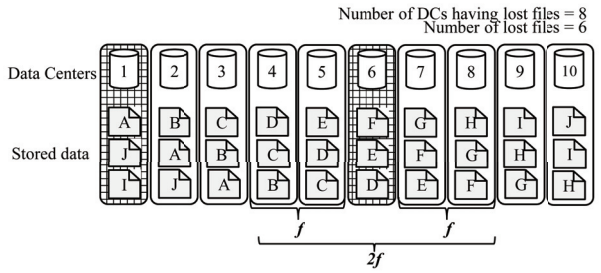
When p DCs are damaged, remaining DCs which have the lost file transmit the lost file to MDRUs. It is clear that the number of lost file and remaining DCs which have the lost file depends on the values of f and p. In addition to this, the number of lost file and remaining DCs with have the lost file are also affected by the combination of damaged DCs as shown in Fig. 3.

In this paper, we analyze combination of malfunctioning DCs, which result in the highest number of lost files as shown in Fig 3(b). Since the file transmission scheduling guarantees that each DC will get to transmit the same amount of lost files, it is possible to simply just calculate the amount of transmission of a single DC. The amount of transmission is $mp(f+1)$ if $(f+1) \geq (N_{\rm dc} - p)$. Here, $(N_{\rm dc} - p)$ is the number of remaining DCs. When $2fp < (N_{\rm dc} - p)$ where $2f$ is the number of DCs that have to send the lost files, the remaining DCs which have to transmit the lost files transmit at most $\lceil \frac{mp(f+1)}{2fp} \rceil$ files. When $2fp \geq (N_{\rm dc} - p)$, all of remaining DCs transmit at most $\lceil \frac{mp(f+1)}{(N_{\rm dc}-p)} \rceil$ files. When $(f+1)p \geq N_{\rm dc}$, each remaining DC transmits at most $\lceil \frac{N_{\rm file}}{(N_{\rm dc}-p)} \rceil$ files because all of files $N_{\rm file}$ are transmitted by all of remaining DCs. Each DC which stores the lost files transmits the files to the satellite. Here, supposed that the propagation delay is ignored because it is identical for each DC, the transmission time which is required to restore the lost files can be calculated by amount of transmission from each DC that sends the lost files over the uplink capacity between the satellite and DCs. Therefore, the transmission time on uplink communication, $\Upsilon_{\rm rep}$, can be formulated as follows.

$$\Upsilon_{\rm rep} \quad = \quad \begin{cases} \frac{\overline{s}}{c}\lceil \frac{N_{\rm file}}{N_{\rm dc}-p} \rceil, & \text{if} \quad (f+1)p \geq N_{\rm dc}, \\ \frac{\overline{s}}{c}\lceil \frac{m(f+1)p}{N_{\rm dc}-p} \rceil, & \text{elseif} \quad 2fp \geq N_{\rm dc} - p, \\ \frac{\overline{s}}{c}\lceil \frac{m(f+1)}{2f} \rceil, & \text{otherwise.} \end{cases} \tag{2}$$



(a) Lowest total transmission volume



(b) Highest total transmission volume

Fig. 3. Impact of different sets of malfunctioning DCs on total transmission volume when $p = 2$ and $f = 2$.

The satellite needs to transmits $m(f+1)$ files to each MDRU because malfunctioning DC stores $(f + 1)$ files. Thus, the transmission time on downlink communication, $\Delta_{\rm rep}$, can be formulated as follows.

$$\Delta_{\rm rep} = \frac{\overline{s}m(f+1)}{c} = \frac{S_{\rm rep}}{c}. \quad (3)$$

We can derive the transmission time from DCs to MDRUs which is required to restore the lost files, $T_{\rm rep}$. In comparison between transmission time on uplink and downlink communication, the larger one becomes $T_{\rm rep}$. Thus, $T_{\rm rep}$, can be formulated as follows.

$$T_{\rm rep} = \max\{\Upsilon_{\rm rep}, \Delta_{\rm rep}\},$$

here, $\Upsilon_{\rm rep} \leq \Delta_{\rm rep}$

$$T_{\rm rep} = \Delta_{\rm rep}. \quad (4)$$

C. Erasure coding

Fig 2(b) shows the data distribution procedure to restore the lost data when f DCs are damaged in the integrated storage system with erasure coding [10]. First, each file is divided into k fragments. Then, $f(= N_{\rm dc} - k)$ parity fragments are made from k fragments by using erasure coding. Subsequently, all fragments are distributed to $N_{\rm dc}$ DCs. Such procedure continues until all files are distributed, namely $N_{\rm file}$ times. While the number of files varies depending on the value of f in the system with replication contrast, we notice that the size of files varies in the system with erasure coding. The storage volume on each DC, $S_{\rm era}$, is defined by the following equation because the fragment size of each file is equal to $\overline{s}/k$ as shown in Fig. 2(b).

$$S_{\rm era} = \frac{\overline{s}N_{\rm file}}{k} = \frac{\overline{s}N_{\rm file}}{(N_{\rm dc} - f)}. \quad (5)$$

When p DCs are damaged by disasters, $(N_{\rm dc}-p)$ remaining DCs transmit $N_{\rm file}(N_{\rm dc} - f)$ fragments. Thus, the number of transmission fragment on each remaining DC becomes at most $\lceil N_{\rm file}(N_{\rm dc} - f)/(N_{\rm dc} - p)\rceil$. Therefore, the transmission time on uplink communication, $\Upsilon_{\rm era}$, can be formulated as follows.

$$\Upsilon_{\rm era} = \frac{\overline{s}}{(N_{\rm dc} - f)c}\lceil\frac{N_{\rm file}(N_{\rm dc} - f)}{(N_{\rm dc} - p)}\rceil. \quad (6)$$

$N_{\rm file}(N_{\rm dc} - f)$ fragments are necessary to decode the lost fragment on each DC in the system with erasure coding. Thus, in downlink communication, the satellite needs to transmit $N_{\rm file}(N_{\rm dc} - f)$ fragments to each MDRU. Thus, the transmission time on downlink communication, $\Delta_{\rm era}$, can be formulated as follows.

$$\Delta_{\rm era} = \frac{\overline{s}}{(N_{\rm dc} - f)c}N_{\rm file}(N_{\rm dc} - f) = \frac{\overline{s}N_{\rm file}}{c}. \quad (7)$$

The transmission time on the downlink communication is longer than that of the uplink communication. Thus, the transmission time from DCs to MDRUs which is required to restore the lost fragments, $T_{\rm era}$, can be formulated as,

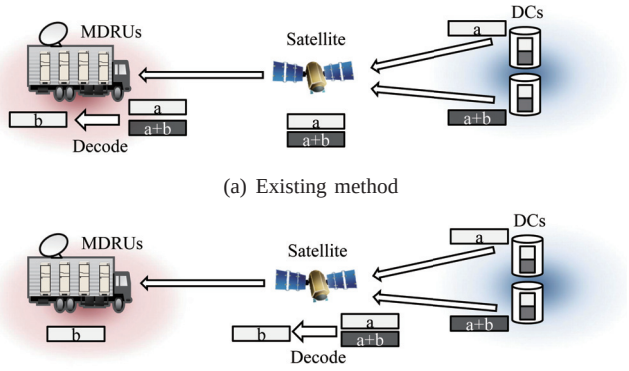
$$T_{\rm era} = \Delta_{\rm era}. \quad (8)$$

In comparison with the replication, the erasure coding takes longer time to transfer the data. We consider a method to reduce the transmission time in the system with erasure coding, because it is necessary to rapidly restore the lost data in disaster areas.

III. Proposed Method

We propose a method to reduce the transmission time in the integrated storage system with erasure coding. In comparison with uplink communication, downlink communication takes longer time to transfer the lost data. Therefore, we focus on network coding technique and attempt to shorten the transmission time on downlink communication by reducing the amount of data which are transmitted from a satellite to MDRUs.

In existing system that utilizes erasure coding, MDRUs perform the decoding process to restore the lost fragments as shown in Fig. 4(a). Thus, the satellite transmits $N_{\rm file}k$ fragments to each MDRU. Consequently, this method is inefficient and takes longer time to transfer the fragments. In contrast to the existing method, we propose a novel method that performs decoding process at the satellite instead of MDRUs as shown in Fig. 4(b). The proposed method results in a significant reduction in the number of fragments which are transmitted from the satellite to MDRUs, and its value becomes $N_{\rm file}$. Unfortunately, the satellite needs to wait until it receives a set of fragments of a file in order to decode and transmit the decoded fragments. Consequently, the wait time is added to the transmission time on downlink communication. Thus, the transmission time of the proposed method, $T_{\rm era}^{\rm pro}$, can be



(a) Existing method

(b) Proposed method

Fig. 4. Comparison of existing method and proposed method.

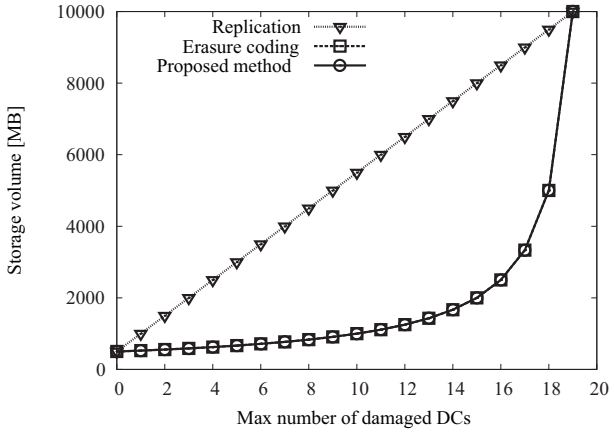


Fig. 5. Impact of maximum number of damaged DCs on storage volume of each data center.

formulated as following equations.

$$T_{\mathrm{era}}^{\mathrm{pro}} = \max\{\frac{\Upsilon_{\mathrm{era}}}{c}, \frac{\overline{s}N_{\mathrm{file}}}{(N_{\mathrm{dc}}-f)c} + \frac{\overline{s}}{(N_{\mathrm{dc}}-f)c}\} = \max\{\frac{\Upsilon_{\mathrm{era}}}{c}, \frac{\overline{s}(N_{\mathrm{file}}+1)}{(N_{\mathrm{dc}}-f)c}\}. \quad (9)$$

IV. Performance Evaluation

In this section, we aim to verify the performance of the proposed method compared with the system based on erasure coding and replication. Especially, we evaluate the communications performance such as the storage volume and the transmission time on uplink (and downlink) communication when the actual number of damaged DCs p and the maximum number of damaged DCs f change.

A. Parameter settings

The parameters, which defines the considered distributed storage system, is summarized in Table II. The number of satellites and DCs are set to 1 and 20, respectively. Suppose that the bandwidth allocation method is FDMA (and PAMA) based system where the size of guard band is ignored, while uplink and downlink capacity are set to 1 MB/s. Moreover, the system distributes $10,000$ files each with 1 MB of data to all DCs.

B. Numerical results

First, we investigate the relationship between the maximum number of damaged DCs and the storage volume on each DC, which is necessary to restore the lost data. Fig. 5 shows the storage volume on each DC when the maximum number of damaged DCs changes from 0 to 19. It is clear from Fig. 5 that the storage volume in the system with proposed method and erasure coding is lower than that of the replication while the storage volume of each DC increases with the increase of the maximum number of damaged DCs. The storage volume in the system with the proposed method is equal to that of erasure coding because data distribution method of the proposed method is the same as that of erasure coding. We can verify that the proposed method and erasure coding are the suitable data distribution methods when the storage volume is taken into consideration.

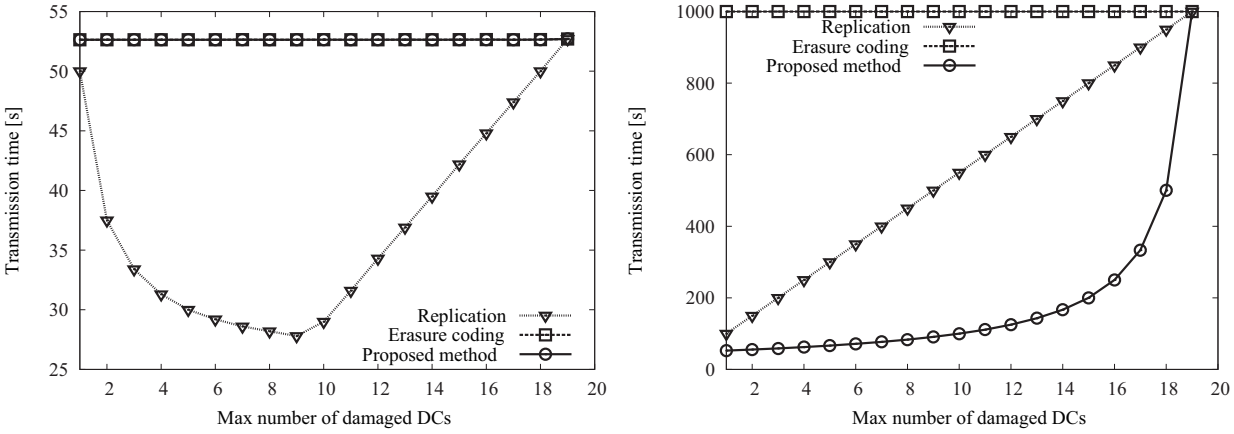
Next, we evaluate the relationship between the maximum number of damaged DCs and the transmission time to restore the lost data. Fig. 6 and Fig. 7 show the transmission time when the actual number of damaged DCs is 1 and 2, respectively. In the case of uplink communication, the performance of the proposed method is same as erasure coding because the proposed method focus on only downlink communication. The system with replication achieves the lowest value when $p = 1$ as shown in Fig. 6(a). However, when the actual number of damaged DCs increases, i.e., $p = 2$, the performance of replication becomes unstable as shown in Fig. 7(a). The comparison of results between uplink and downlink communication clearly demonstrates that the downlink communication takes longer time in any case. We can notice that the transmission time on downlink communication have more impact on system performance. In the case of downlink, while the system with erasure coding takes the longest time to transfer data, the transmission time of the proposed method is the lowest by decoding data at satellite and reduction of the transmission volume. Therefore, we can verify that proposed method is the suitable data distribution method also when the transmission time is taken into consideration.

TABLE II
Parameter settings.

Parameter	Value
Number of satellites	1
Number of DCs N_{dc}	20
Up/downlink capacity C	10 MB/s
File size $\overline{s}$	1 MB
Number of files N_{file}	10000

V. Conclusion

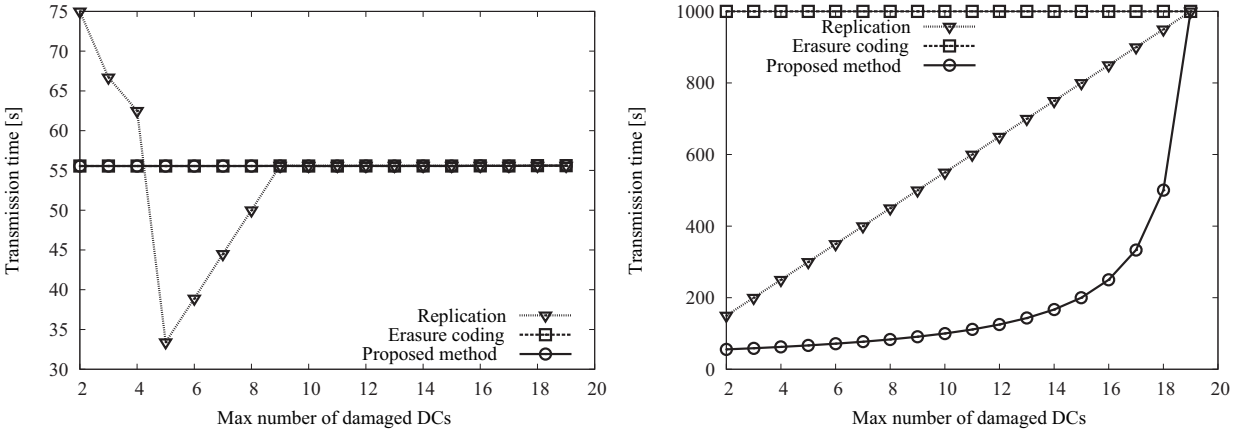
In this paper, we have investigated the impact of data distribution method on communications performance in the distributed storage system integrated satellite networks, DCs,



(a) Transmission time on uplink communication

(b) Transmission time on downlink communication

Fig. 6. Impact of maximum number of damaged DCs on transmission time when actual number of damaged DCs $p = 1$.



(a) Transmission time on uplink communication

(b) Transmission time on downlink communication

Fig. 7. Impact of maximum number of damaged DCs on transmission time when actual number of damaged DCs $p = 2$.

and MDRUs. The numerical analysis demonstrates that the storage volume of the erasure coding becomes lower than that of the replication while the transmission time becomes higher. Therefore, in order to shorten the transmission time, we have proposed a novel method that uses the satellite to decode the data instead of MDRUs and transmits the decoded data to MDRUs. Through numerical results, we have verified that the proposed method achieves a higher performance.

However, the performance of the proposed method decreases when the maximum number of damaged DCs increases. It would be interesting to further extend the efficient transfer data method and explore the suitable maximum number of damaged DCs. We also want to consider another traffic scenario, where the enable link capacity dynamically changes.

REFERENCES

[1] Y. Wu, " Existence and Construction of Capacity-Achieving Network Codes for Distributed Storage, "*IEEE Journal on Selected Areas in Communications*, Vol. 28, No. 2, pp. 277-288, Feb. 2010.

[2] A. G. Dimakis, P.B. Godfrey, Y. Wu, M. J. Wainwright, K. Ramchandran," Network Coding for Distributed Storage Systems, "*IEEE Transactions on Information Theory*, Vol. 56, No. 9, pp. 4539-4551, Sept. 2010.

[3] Ministry of Internal Affairs and Communications, " 2011 WHITE PAPER Information and Communications in Japan, "Part 1, 2011.

[4] SKY Perfect JSAT Corporation: S*Plex3, available at, http://www.splex3.com/

[5] Y. Kawamoto, H. Nishiyama, N. Kato, N. Yoshimura, and N. Kadowaki, " A delay-based trafc distribution technique for Multi-Layered Satellite Networks, " *In Proc. of Wireless Communications and Networking Conference (WCNC)*, pp. 2401-2405, Apr. 2012.

[6] A. G. Dimakis, V. Prabhakaran, K. Ramchandran " Decentralized Erasure Codes for Distributed Networked Storage, "*IEEE Transactions on Information Theory*, Vol. 52, No. 6, pp. 2809-2816, Jun. 2006.

[7] T. Sakano and A. Takahara, " A resilient network architecture taking account of the ICT demand characteristics in large scale disasters , " *Technical Report of IEICE*, vol. 112, no. 118, pp. 73-78, Jul. 2012.

[8] W. Liu, H. Nishiyama, N. Kato, Y. Shimizu, and T. Kumagai " A Novel Gateway Selection Method to Maximize the System Throughput of Wireless Mesh Network Deployed in Disaster Areas, " *In Proc. of International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 787-792, Sep. 2012.

[9] C. Gkantsidis and P. R. Rodriguez, " Network coding for large scale content distribution, "*In Proc. of Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCM)*, vol. 4 pp. 2235-2245, Mar. 2005.

[10] H. Weatherspoon and J. D. Kubiatowicz, " Erasure Coding vs. Replication: A Quantitative comparison, "*Revised Papers from International Workshop on Peer-to-Peer Systems*, pp. 328-338, Mar. 2002.