# Probability-Symmetric Storage Allocation for Distributed Storage Systems based on Network Coding

http://dx.doi.org/10.3991/ijoe.v9iS4.2659

Qingtao Wu, Xulong Zhang, Ruijuan Zheng and Mingchuan Zhang
Henan University of Science and Technology, Luoyang, P. R. China.

***Abstract*—The goal of optimal allocation is to increase stored data availability subject to minimizing the storage budget. Symmetric allocation based on the network coding has been proved to be optimal for distributed storage systems if node availability is not considered. However, because of network conditions and the inherent properties of nodes, each node will have a different availability. This paper focuses on the problem of optimizing distributed data storage when considering node availability. Using a probability model for storage systems, we redefine symmetric allocation in terms of probability-symmetric allocation and propose a probability-symmetric allocation model and strategy based on network coding. These are shown to be optimal in general. Compared with the symmetric allocation scheme proposed by Leong et al., the proposed probability-symmetric allocation scheme not only improves data availability in distributed storage systems but also is more practical.**

***Index Terms*—Distributed storage allocation; Network coding; Probability-symmetric**

## I. INTRODUCTION

For sensor networks and networked data centers, distributed data storage security and efficiency are the focus of much current research. Traditional data backup methods will cost more, and traditional data-synchronization methods are not reliable enough for distributed network storage systems. Network coding [1] offers distributed data storage with high reliability and at low cost. For the distributed data storage described in [2], which is based on network coding, the original data is expanded into a plurality of encoded data blocks, with the coded data blocks allocated to different data storage nodes. To access the original data, data is received from data storage nodes containing encoded data blocks, and restored by calculation [3]. An allocation strategy for the coded data blocks is the key problem in distributed data storage based on network coding.

To solve the allocation problem for coded data blocks, Leong and Dimakis [4] proposed a uniform code-block allocation model, for which the encoding of the data block amounts assigned to each storage node and the model were proved to be optimal. This ensured the availability of data with a minimal data-allocation overhead. In a further study, Leong et al. set the same failure rate for all data storage nodes, namely that the data-receiving node had the same probability of accessing a data storage node to retrieve a data block. At this time, the uniform allocation strategy remains the optimal allocation strategy. This is because each storage node itself has a certain failure rate, and storage nodes of different types, ages, and failure rates make different use of the environment. In considering the node availability of Leong, the uniform code-block allocation model proposed by Dimakis does not consider the availability of storage nodes, and the same node failure rate, rather than the optimal mode.

In reality, inherent properties of the storage-node network environment and of the nodes themselves determine that the data receivers accessing storage nodes will have a certain probability of failure. Examples of failure include network communication link failure, a node's online storage duration, and hacker attacks. All these may prevent the data receiver from accessing data storage nodes. At this time, data storage nodes with a certain probability of failure are assumed to determine the node availability for data receivers.

This paper focuses on the distributed data storage optimization when considering storage node availability. Using methods from probability theory, we establish a storage-node probability distribution based on a data model. Based on a redefined uniform-distribution model and the probability distribution, we present a data-storage probability distribution strategy and method, and demonstrate that the proposed method is optimal. Compared with Leong and Dimakis, who proposed a uniform distribution model, the model and the method proposed in this paper consider the node failure rate, thereby improving the availability of data storage and modeling the distributed storage system more realistically.

## II. RELATED WORK

In distributed storage systems, network coding is often used to improve data availability and reliability. Current academic research discusses how best to utilize data redundancy to improve reliability. Approaches to data-redundancy methods include simple backup, erasure coding, and network coding. Compared with simple backup, erasure codes can provide higher data reliability for the same storage overhead [5]. At the same time, network coding can be applied to distributed storage, thereby balancing storage space and bandwidth. Distributed data storage based on network coding is a potential way forward.

For simple backup, the data source needs only to fully replicate the original data stored in the data storage node. For distributed storage based on network coding, where the data storage nodes require that data blocks receive

coded data, the storage space and use of data bandwidth are better than for simple backup [3]. Based on maximum distance separable (MDS)codes, particularly via the MDS(n, K) distributed data storage method, this coding can effectively reduce redundant data; balance the storage space and network bandwidth; reduce the data distribution, storage, and access requirements; and improve reliability [6].

Distributed storage methods based on network coding have received widespread attention in academic circles [1][2][3]. However, research into the distribution of code blocks is less common. The traditional method [4][7] is to assign one code block per storage node.

In fact, for distributed data storage based on availability, knowledge of the network storage-node topology can improve the efficiency and usability of the block allocation for coded data. Research work in this area is still relatively limited, mainly focusing on two aspects: (1) data distribution methods based on efficiency; (2) data allocation methods based on usability.

Among data distribution methods based on efficiency, the minimum-storage regenerating method, using minimum-memory regeneration codes, aims to reduce the data storage space and improve the repairing efficiency, with data-coding modes of distribution as the target. The minimum-bandwidth regenerating method using minimum-bandwidth regeneration code considers channel bandwidth, using data coding to reduce the link bandwidth allocation for the target [8][9]. Tree-type data allocation methods have been proposed [10], for which a data allocation strategy is presented from the perspective of bandwidth, aiming to improve the repair efficiency for data.

For data distribution methods based on availability, Leong and Dimakis [4][7] give a definition of uniform distribution and propose a storage strategy for distributed storage systems with uniform distribution. The strategy is based on an ideal model, which does not consider the storage space, storage-node failure probability, or the number of storage nodes. For this ideal model, the authors prove that a uniform distribution is optimal. Data-receiving nodes can access all data storage nodes for the data block to ensure that restoration of the original data is completely successful. In considering the efficiency of working nodes, Leong [7] proposed a data distribution method using probability measures. This method considers the probability of failure for all storage nodes to be the same. The essence of this method is consistent with the methods in the literature [4]. The author proves that the method is optimal only if the premise of the same node-failure probability is met.

The advantage of these two methods is that the storage allocation method is simple. The main disadvantage is the lack of consideration of data-distribution node characteristics and the network environment, thereby not meeting the needs of practical applications.

In reality, individual storage node characteristics determine whether users can access data storage nodes, not all of which have the same failure probability, and the present study did not pay attention to the data storage nodes. Different probabilities of failure for different storage nodes will directly affect the reliability of recovering the original data, and a uniform distribution strategy will no longer be the optimal allocation strategy. System reliability is an important aspect of distributed storage systems. We focus on a different storage-node data distribution method for allocating coded data from general blocks by considering the different failure probabilities of storage nodes in the data distribution network.

## III. PROBABILITY DISTRIBUTION MODEL BASED ON NETWORK CODING

In distributed storage systems, storage nodes will have different failure probabilities. Therefore, we cannot assume that the traditional uniform data distribution method will be optimal. A more feasible data distribution method would assign different numbers of data blocks to data storage nodes according to the different node availabilities. Based on considerations of storage node availability, a probability distribution for network coding can be developed from the uniform distribution model.

### *A. Establish the hypothesis and model*

The distributed network storage system studied in this paper has one data source node and M data storage and data-receiving nodes, as shown in Figure 1.

We use $MDS(n,k)$ encoding [2][3], encoding the original data into coded data blocks. The original data unit is divided into k blocks $F_1,\mathrm{L}\ ,F_k$ to become an n-block coding-matrix code, $\mathrm{B}_1,\mathrm{L}\ ,B_n$. Each block Bi is a linear combination of $F_1,\mathrm{L}\ ,F_k$, using the coefficient vector $(a_{i1},\mathrm{L}\ ,a_{ik})^T$, giving:

$$\begin{pmatrix} a_{11} & \mathrm{L} & a_{1k} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ a_{n1} & \mathrm{L} & a_{nk} \end{pmatrix} \cdot \begin{pmatrix} F_1 \\ \mathrm{M} \\ F_k \end{pmatrix} = \begin{pmatrix} B_1 \\ \mathrm{M} \\ B_n \end{pmatrix}.$$

Data allocation assigns the Bi to different data-storage nodes. The code allocation scheme is shown in Figure 1.

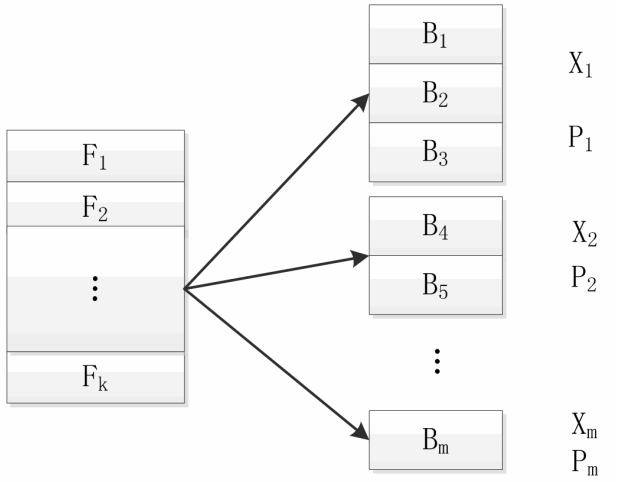


Figure1. Data-allocation coding based on MDS(n ,k)

The source data object D is the data to be stored. Via the $MDS(n,k)$ coding method, a coded data block is

allocated to each storage node represented as $x_i$, $1 \le i \le m$. The availability of a storage node is the probability of the node being available, which will depend on factors such as the manufacturer, the technology, the operation time, and the operational environment. This information can be obtained by node sampling. The availability of each storage node will be different, but over a short period, it may be considered fixed. If the probability of failure for data storage node $xi$ is $F_{xi}$, the availability probability is $p_i = 1 - F_{xi}$.

### B. Uniform distribution of probability model

For data distribution, this paper proposes to use anode-availability strategy based on the proposed probability distribution model. This model is a generalization of the traditional uniform distribution, whereby the distributed storage system considers the node failure rate. The allocation model is as follows:

$$\begin{aligned} X &= \{x_i\}_{i=1}^{m} \\ &= \{x_1, \mathrm{K}, x_m\} \\ &= \{f(p_1), \mathrm{K}, f(p_m)\} \\ \sum_{i=1}^{m} x_i &= \sum_{i=1}^{m} f(p_i) = T \end{aligned}$$

$$f(p_i) > f(p_j), \forall p_i, p_j \in [0,1], p_i > p_j$$

Here, we use $X = \{x_i\}_{i=1}^{m}$ to represent the data distribution, the distribution function $f(p_i)$ is assigned to the i[th] encoded data node, $f(p_i)$ is defined over the interval [0, 1] as a one-way increasing function, changing the availability of node $p_i$. $T$ is the storage space required for the data storage nodes.

With the use of the data distribution model proposed in this paper, allocation of data blocks to high-availability nodes occurs more often, enabling a data receiver access to blocks with higher reliability, and with a higher probability of recovering the original data object.

### C. Data recovery method

For the data-receiving node to successfully recover the original data, access to the data storage node acquiring the data should not be less than for the original data. If the original data is of unit size, the data-receiving node should access a data quantity of not less than 1.

If the data-receiving node is to successfully restore the original data, we need to access a subset *r* of data-storage nodes, using $|r|$ to represent the *R* subset of the number of elements, in terms of storage-node numbers. If *S*is defined as the successful recovery of data events, then:

$$P(S) = \sum_{|r|=1}^{m} P(|r|) \cdot I\left[\sum_{i \in r} x_i \ge 1\right].$$

Among these,

$$P(|r|) = \prod_{i \in r} p_i \prod_{j \in M \setminus r} (1 - p_j)$$

is true if $I[G] = 1$, and G is false if $I[G] = 0$.

Here, the distributed storage system of *m* data storage nodes is composed of a plurality of *r* subsets, where *M\r* represents an *r* subset of the *M* set. This set is defined as the *r* subset of all successful repairs $S_r$, where $|S_r|$ is the number of elements in the collection of successful repairs that is composed of *m* data nodes, and *r* is the number of subsets.

## IV. ANALYSIS AND EVALUATION

This section uses storage-node availability to analyze the probability distribution model, probabilities for a data-receiving node to achieve successful recovery, and data availability. At the same time, this article considers the presence of node failures, data availability, uniform distribution, and probability distribution evaluation.

### A. Data availability

For distributed memory systems in the presence of node availability, data storage nodes exhibit failure rates, implying that the data-receiving node successfully recovering the probability of the original data objects is not the only value. For distributed data storage, the reliability of the system as a distributed system is the most important goal.

For a distributed storage system using the probability distribution model, the amount of data stored in the *m* data storage nodes is different, in accordance with the node availability. If this changes, therefore, the number of subsets of elements in a successful recovery will be uncertain, i.e., $|r| \in [1, m]$, provided the subset of the nodes can supply the quantity of data $\sum_{i \in r} x_i \ge 1$. In distributed storage systems, the data availability for each subset in a successful recovery and the data availability of each *r* subset of nodes are available.

Theorem 1: with the possibility of storage-node failures in a distributed storage system, the data availability (or probability of data recovery) is:

$$\sum_{|r|=1}^{m} \min\left(\frac{|r|T}{m}, 1\right) P(|r|).$$

Proof: for the data-receiving node to successfully restore the original data, the data size must meet the condition:

$$\sum_{i \in r} x_i = \sum_{i \in r} f(p_i) \ge 1.$$

Let $S_r$ denote a successful subset of the set of all $r$. $|S_r|$ is the number of successful elements in the collection. $S$ is defined as the successful recovery of data events, and $A_i$ denotes access to exactly *i* nodes. Then:

$$P(S)=\sum_{|r|=1}^{m}P(S|A_i)\cdot P(A_i)$$
$$=\sum_{|r|=1}^{m}\frac{|S_r|}{\binom{m}{|r|}}\cdot\prod_{i\in r}p_i\prod_{j\in M\setminus r}(1-p_j).$$
$$=\sum_{|r|=1}^{m}\frac{|S_r|}{\binom{m}{|r|}}\cdot P(|r|)$$

Each subset of $r$ can be represented by the equation:

$$\sum_{i\in r}x_i=\underbrace{x_{t1}+x_{t2}+\mathrm{L}+x_{tr}}_{r\quad nodes}.$$

Each element in $S_r$ represents successful recovery of a subset of $S_r$, leading to the following equation:

$$\begin{cases}x_{t1}+x_{t2}+\mathrm{L}+x_{tr}\ge 1\\ \qquad\mathrm{M}\\ x_{S_r1}+x_{S_r2}+\mathrm{L}+x_{S_rr}\ge 1\end{cases}.$$

For these equation, $a_1x_1+\mathrm{L}+a_mx_m\ge|S_r|$ because each storage node belongs to a different subset of $R$, $0\le a_i\le\binom{m-1}{|r|-1}$. Therefore:

$$|\mathrm{S}_r|\le a_1x_1+\mathrm{L}+a_mx_m$$
$$\le\binom{m-1}{|r|-1}\sum_{i=1}^{m}x_i$$
$$\le\binom{m-1}{|r|-1}T$$

because $|S_r|\le\binom{m}{|r|}$ is $|S_r|\le\min\left(\binom{m-1}{|r|-1}T,\binom{m}{|r|}\right)$.

Therefore, the reliability of node availability in the distributed storage system is:

$$P(S)=\sum_{|r|=1}^{m}\min\left(\frac{|r|T}{m},1\right)P(|r|).$$

The system reliability in distributed storage systems, namely the probability of recovery success for the data-receiving node, is determined by a random variable $p_i$, $i = 1,..., m$.

### *B. Probability distribution model and distribution model comparison*

For a distributed storage system, the data distribution strategy is the main factor affecting the performance of the system. Dimakis has proposed a uniform allocation strategy, which is the optimal allocation strategy for an ideal model. Using the distributed storage system's storage-node availability, the probability distribution model proposed in this paper considers the storage-node failure rate, and the distribution function for the allocation to storage nodes does not involve equal amounts of data. If all storage nodes are available, our model is equivalent to the uniform distribution model in [4]. If all storage-node availabilities are equal, our model is equivalent to that in [7]. However, our uniform probability model has more universal application than the uniform distribution model.

Theorem 2: by considering node availability in distributed-storage systems, the uniform probability model is superior to [4] and [7] in the literature. To prove Theorem 2, we first need a lemma.

Lemma 1: the minimum number of nodes in the probability distribution model for successful recovery is less than that for the uniform distribution defined in [4] and [7].

Proof: we use a reductio-ad-absurdum proof.

The minimum number of nodes to evenly distribute the successful recovery of the assumed probability across the data-storage nodes, using $r_1$ in the appropriate distribution function, is: $f(p_1)>f(p_2)>\mathrm{L}>f(p_{r1})$.

For the minimum number of nodes to evenly distribute the successful recovery using $r_2$, the data-storage node weight is $\frac{T}{m}$, where $r_1>r_2$. The number of successful recovery nodes with the least at the receiving node, having data access equal to 1, is:

$$f(p_1)+f(p_2)+\mathrm{L}+f(p_{r1})=r_2\cdot\frac{T}{m}=1$$
$$\Rightarrow f(p_1)-\frac{T}{m}+f(p_2)-\frac{T}{m}+\mathrm{L}+\sum_{i=r_1-r_2+1}^{r_1}f(p_i)=0$$

Because $f(p_i)>\frac{T}{m}$ and $\sum_{i=r_1-r_2+1}^{r_1}f(p_i)>0$, this cannot be equal to zero, thereby contradicting the assumption. We therefore have $r_1<r_2$ as the minimum number of nodes. The uniform distribution model successfully restored the data with a probability less than successful recovery using a uniform distribution.

We now prove Theorem 2.

Proof: in the uniform distribution model, all data storage nodes store the same amount of data. Without considering empty nodes, the amount of data stored in all storage nodes is $\frac{T}{m}$.In an ideal situation, successful restoration requires at least $\left\lceil 1/\frac{T}{m}\right\rceil=\left\lceil\frac{m}{T}\right\rceil$ nodes. At this time, $\frac{|r|T}{m}>1$ in existing distributed storage

systems. Node availability in the uniform distribution model of data availability is $\sum_{|r|=\left\lceil \frac{m}{T} \right\rceil}^{m} P(|r|)$.

The minimum number of nodes in the probability distribution model for successful restoration is $t$, where $t < \left\lceil \frac{m}{T} \right\rceil$. The difference between this availability of data and data availability for the uniform distribution model is:

$$\sum_{|r|=t}^{m} \min\left(\frac{|r|T}{m}, 1\right) P(|r|) - \sum_{|r|=\left\lceil \frac{m}{T} \right\rceil}^{m} P(|r|)$$

$$= \begin{cases} \sum_{|r|=t}^{\left\lceil \frac{m}{T} \right\rceil - 1} P(|r|) & \frac{|r|T}{m} > 1 \\ \sum_{|r|=t}^{m} \frac{|r|T}{m} P(|r|) - \sum_{|r|=\left\lceil \frac{m}{T} \right\rceil}^{m} P(|r|) & \frac{|r|T}{m} < 1 \end{cases}$$

When $\frac{|r|T}{m} > 1$, $\sum_{|r|=t}^{\left\lceil \frac{m}{T} \right\rceil - 1} P(|r|) > 0$.

For the probability distribution, data availability is greater than for the uniform distribution.

When $\frac{|r|T}{m} < 1$, the uniform distribution model fails to restore the original data. To sum up, the probability distribution model of data availability is more effective than the uniform distribution of data availability. If there is node availability information for the distributed storage system, the probability distribution strategy is better than the uniform allocation strategy.

## V. CONCLUSIONS

Optimization of distributed data storage has the goal of ensuring safe data recovery from storage to improve the reliability of data storage systems. Without considering storage-node availability, a uniform distribution based on network coding has been shown to be optimal. However, different storage nodes can have different availability, caused by node failure and other factors. Based on probabilities for storage-node availability instead of a uniform distribution model, we propose a probability distribution strategy and method and demonstrate that the proposed method is optimal. In contrast to Leong et al., Dimakis proposed a uniform distribution model, following which the model and method proposed in this paper consider the availability of nodes, thereby improving the effectiveness of the data storage system and being more realistic in practice.

## REFERENCES

[1] A. G. DIMAKIS, V. PRABHAKARAN, and K. RAMCHANDRAN, "Decentralized Erasure Codes for Distributed Networked Storage," IEEE Transactions on Information Theory, vol. 52(6), pp. 2809-2816, 2006. http://dx.doi.org/10.1109/TIT.2006.874535

[2] A. G. DIMAKIS, K. RAMCHANDRAN, and Y. WU, "A Survey on Network Codes for Distributed Storage," Proceedings of the IEEE, vol. 99(3), pp. 476-489, 2011. http://dx.doi.org/10.1109/JPROC.2010.2096170

[3] A. G. DIMAKIS, P. B. GODFREY, and Y. WU "Network Coding for Distributed Storage Systems," IEEE Transactions on Information Theory, vol. 56(9), pp. 4539-4551, 2010. http://dx.doi.org/10.1109/TIT.2010.2054295

[4] D. LEONG, A. G. DIMAKIS, and T. HO "Distributed storage allocation problems," Workshop on NetCod '09, Lausanne, 2009.

[5] S. ACEDANSKI, S. DEB, and M. MEDARD "How good is random linear coding based distributed networked storage," in Proc. 1st Workshop on Network Coding, Riva del Garda, Italy, Apr. 2005.

[6] V. R. CADAMBE, S. A. JAFAR, and H. MALEKI "Minimum Repair Bandwidth for Exact Regeneration in Distributed Storage," Wireless Network Coding Conference, 2010 IEEE, Boston, MA, USA,2010.

[7] D. LEONG , A. G. DIMAKIS, T. HO "Distributed Storage Allocation for High Reliability," 2010 IEEE International Conference on Communications (ICC), Cape Town, South Africa, 2010. http://dx.doi.org/10.1109/ICC.2010.5502492

[8] K. V. RASHMI, N. B. SHAH, and P. V. KUMAR "Optimal Exact-Regenerating Codes for Distributed Storage at the MSR and MBR Points via a Product-Matrix Construction," IEEE Transactions on Information Theory, vol. 8(57), pp. 5227-5239, 2011. http://dx.doi.org/10.1109/TIT.2011.2159049

[9] K. V. RASHMI, N. B. SHAH, and P. V. KUMAR "Explicit and Optimal Exact-Regenerating Codes for the Minimum-Bandwidth Point in Distributed Storage" 2010 IEEE International Symposium on Information Theory Proceedings (ISIT), Austin, TX,2010. http://dx.doi.org/10.1109/ISIT.2010.5513367

[10] J. LI, S. YANG, and X. WANG "Tree-structured Data Regeneration with Network Coding in Distributed Storage Systems ," 17th International Workshop on Quality of Service, Charleston, 2009.

[11] D. LEONG, A. G. DIMAKIS, and T. HO "Symmetric Allocations for Distributed Storage," Global Telecommunications Conference, Miami, FL, USA, 2010.

## AUTHORS

**Q.T. Wu** is with the Electronic & Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China (e-mail: wqt8921@126.com).

**X. L. Zhang, R.J. Zheng** and **M.C. Zhang** are with the Electronic & Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China. (Email: daleloogn@126.com, rjwo@163.com, zhlzmc@163.com).

This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. U1204614 and No. 61003035, and in part by the Plan for Scientific Innovation Talent of Henan Province under Grant No. 124100510006. This article is an extended and modified version of a paper presented at the International Conference on Mechanical Engineering, Automation and Material Science (MEAMS2012), held 22-23 December 2012, Wuhan, China. Received 6 January 2013. Published as resubmitted by the authors 01 May 2013.