# Distributed Storage Allocations for Optimal Service Rates

Pei Peng, Moslem Noori, *Member, IEEE*, and Emina Soljanin, *Fellow, IEEE*

***Abstract*— Distributed systems operate under storage access and download service uncertainty. We consider two access models. In one, a user can access each storage node with a fixed probability, and in the other, a user can access any fixed-size subset of nodes. We consider two download service models. In the first (small file) model, the time to transmit file data is negligible compared to the overall average download time. In the second (large file) model, the download time scales with the amount of downloaded data. The performance metric is the system's service rate. For a fixed redundancy level, the systems' service rate depends on the allocation of coded chunks over the storage nodes. Since finding the general optimal allocation is prohibitively hard, we consider quasi-uniform allocations, where coded content is equally spread among a subset of nodes. The question we address asks what the size of this subset (spreading) should be. We show that concentrating the coded content to a minimum-size subset is universally optimal for the small file model. However, for the large file model, the optimal spreading depends on the system parameters. These conclusions hold for both access models.**

***Index Terms*— Distributed storage systems, service rate, optimal allocations, erasure coding, redundancy.**

## I. Introduction

DISTRIBUTED storage systems (DSSs) are a vital part of computing and content providing environments, such as cloud data centers and edge systems. Their purpose is to ensure reliable storage and quick access of data by end-users or computing processes. Today, both goals are commonly addressed by storing data redundantly. The DSS performance must be robust to various forms of uncertainty. Most of the current work addresses internal uncertainty (e.g., straggling) in operations of the system itself (see, e.g., [3]–[6]). This paper considers the uncertainty in both network accessibility and download services, which are common in edge computing [7].

Manuscript received February 7, 2021; revised May 31, 2021 and June 9, 2021; accepted June 19, 2021. Date of publication July 9, 2021; date of current version October 18, 2021. Part of this research is based upon work supported by the National Science Foundation under Grant No. CIF-1717314. This article was presented in part at the Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT) [1] and in part at the 2018 56th Annual Allerton Conference on Communication, Control, and Computing [2]. The associate editor coordinating the review of this article and approving it for publication was L. Ong. *(Corresponding author: Pei Peng.)*

Pei Peng and Emina Soljanin are with the Electrical and Computer Engineering Department at Rutgers, The State University of New Jersey, Piscataway, NJ 08854 USA (e-mail: pei.peng@rutgers.edu).

Moslem Noori is with 1QB Information Technologies (1QBit), Vancouver, BC V6E 4B1, Canada (e-mail: moslem.noori@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCOMM.2021.3095968.

Digital Object Identifier 10.1109/TCOMM.2021.3095968

We address the following network access uncertainties, as considered in [8] and the follow-up work. 1) Users may only be able to access a random subset of nodes. Such users can retrieve a file if the storage content of the accessed nodes suffices for file decoding. 2) Even when users have access to all nodes, a node may not respond. Such users can retrieve the file if the storage content of the responding nodes suffices for file decoding. We adopt a storage model of [8], where files are split into chunks, and redundancy is introduced at some fixed level, determined by the storage budget that the DSS has for the file. The total storage is the only constraint. There is no limit on how many chunks a particular node can store.

We consider two download service models. In the first download model, the time it takes to transmit file data is negligible compared with the overall average download time. Thus the download time is a random variable that only depends on the storage system parameters. We refer to this model as the *small file* scenario. In the second download model, there is randomness associated with both the file transmission and inherent system's operations. Thus the download time scales with the amount of data being downloaded. We refer to this model as the *large file* scenario. We adopt two common service and scaling models used in the literature. For more detail and other models, see [9], [10] and references therein.

Two important DSS performance measures have been considered in the literature [1], [2], [8], [11]–[14]. One is the *probability of successful data recovery* and the other is the *average service rate*. Finding these quantities has been challenging, and the optimal allocations are known only in some special cases. Some versions of this problem are related to a long-standing conjecture by Erdős on the maximum number of edges in a uniform hypergraph [15]. In general, both measures are of interest and should be simultaneously taken into account. Increasing the chance of successful download, while desirable, should not come at the cost of intolerable delivery delay. Moreover, in practice, we may want to partially sacrifice a successful but tardy data delivery to some users to ensure that other users, that can receive the data, are served fast. This paper focuses on the service rate of a DSS. Service rate is an emerging increasingly important performance measure, which addresses stability in distributed systems with redundancy under uncertainty in service request arrivals [16]–[22].

For a fixed redundancy level, the system's service rate is determined by the allocation of coded chunks over the storage nodes. We consider quasi-uniform storage allocations, where coded content is uniformly spread among a subset of nodes, and ask what the size of this subset (spreading) should be

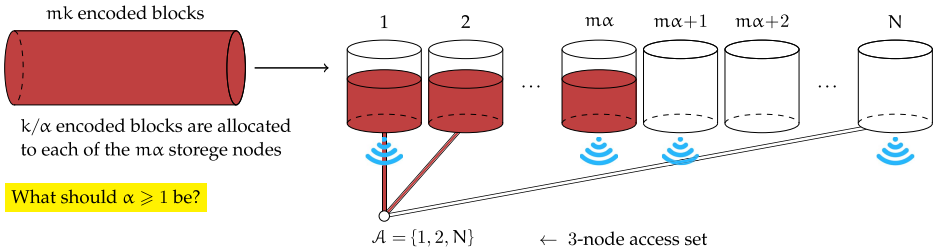



Fig. 1. A DSS with $N$ nodes with quasi-uniform allocation. Each node stores either $k/\alpha$ or 0 data blocks of interest to some users, and thus only $\varphi = m\alpha$ nodes contain data blocks. The WiFi sign indicates that the node is able to serve the user. Note that that is independent of whether or not the node has been accessed or has the data. Here, three nodes are successfully accessed, but only two of them have (coded) data blocks. One of the accessed nodes has data blocks but is not able to serve the user.

to maximize the expected download service rate. We consider two service time models: scaled exponential service time and shifted exponential service time. These and other models are considered in the context of the server-dependent scaling and data-dependent scaling in [9], [10]. We show that concentrating the coded content to a minimum-size subset is universally optimal for the small file model. However, for the large file model, the optimal spreading depends on the system parameters. These conclusions hold for both access models.

The paper is organized as follows: In Sec. II, we present the system architecture and the models for the service time and rate. In Sec. III, we state the problem and summarize the contributions of this paper. In Sec. IV, V, and VI, we characterize the DSS service rate and determine the optimal allocation for three common service time distributions and two different access models. Conclusions are given in Sec. VII.

## II. System Model and Problem Formulation

### A. Storage Model

A file consisting of $k$ blocks is redundantly stored over a DSS with $N$ storage nodes. The file is encoded by a maximum distance separable (MDS) code into $mk$ ($m \in \mathbb{N}$) encoded blocks so that any $k$ of them are sufficient to recover the file. The $mk$ encoded blocks are partitioned into $N$ subsets $\mathcal{S}_i$'s for $i \in \{1, \ldots, N\}$ where $|\mathcal{S}_i| = s_i$, and thus $\sum_{i=1}^{N} s_i = mk$. We refer to such partitioning as *allocation*. The $s_i$ blocks in $\mathcal{S}_i$ are stored at the storage node $i$. Note that $0 \leq s_i \leq k$ since storing more than $k$ blocks on a node is unnecessary.

Optimizing general storage allocations is computationally difficult, see [11]. Thus we focus on quasi-uniform allocations [12], where a node can either store a constant number of blocks $k/\alpha$ ($\alpha \in \mathbb{N}$) or no blocks at all. We will refer to such allocation as $\alpha$ quasi-uniform allocation. Fig. 1 depicts an example of $\alpha$ quasi-uniform allocation on $N$ nodes. We refer to a quasi-uniform allocation where $\alpha = 1$ as *the minimal spreading* allocation [11]. Note that for the minimal spreading allocation, the $k$ file blocks are simply replicated over some $m$ storage nodes. Similarly, an allocation with $\alpha = N/m$ is referred to as a *the maximal spreading* allocation since the file chunks are spread over all $N$ nodes in the system.

### B. Data Access and Delivery Models

Fixed-size Access: In this model, the download request is forwarded to a random $r$-node subset of the $N$ storage nodes. Therefore, since the data is MDS encoded, the access to a given $r$-subset $\mathcal{A}$ results in the successful recovery of the data iff the nodes in $\mathcal{A}$ jointly contain at least $k$ coded blocks:

$$\sum_{i \in \mathcal{A}} s_i \geq k. \tag{1}$$

Probabilistic Access: Here, each download request is forwarded to all $N$ nodes. However, a node does not respond with probability $p$. Let $\mathcal{A}$ be the set of nodes that are successfully accessed. Then the condition for data recovery is again (1). In this case, $1 \leq \alpha \leq \frac{N}{m}$. In this access model, $|\mathcal{A}|$ is a binomial random variable distributed as $\mathrm{Bin}(N, p)$.

Regardless of the access model, for an accessed subset of nodes $\mathcal{A}$, we denote the number of nodes containing data by $\varphi(\mathcal{A})$. Observe that $\varphi(\mathcal{A}) \leq |\mathcal{A}|$. For instance, in Fig. 1, three nodes ($|\mathcal{A}| = 3$) are accessed while only $\varphi(\mathcal{A}) = 2$ of them have data. We assume a request is simultaneously served by all nodes in the accessed set $\mathcal{A}$, where each node takes some i.i.d. random time to deliver its data blocks. In the fixed-size access model, $|\mathcal{A}| = r$, while in the probabilistic access model, $|\mathcal{A}|$ is a Binomial random variable between 1 and $N$. The file can be reconstructed when the accessed nodes jointly deliver $k$ encoded blocks. We assume that a node has to deliver all its blocks for the download to count. For an $\alpha$ quasi-uniform allocation, the download request can be served iff $\varphi(\mathcal{A}) \geq \alpha$, and as soon as all blocks are downloaded from any $\alpha$ out of $\varphi(\mathcal{A})$ nodes. Therefore, the average service time for the file, $T_s(\alpha|\varphi(\mathcal{A}))$, is the expected value of the $\alpha$-th order statistics of $\varphi(\mathcal{A})$ waiting times at the storage nodes.

### C. Download-Time Models

As discussed in the introduction, depending on how the data transmission time compares with the average time the system takes to fulfill the request, we consider the small and large file models. Here, *small* and *large* are informal descriptive terms. The precise mathematical models are stated next.

*1) Small File Model:* When a task is assigned to a storage node, there is some random waiting time before the data transmission starts (needed, e.g., for a general handshake and/or acquiring the requested content). We assume the waiting time follows an exponential distribution. Compared to the mean of this distribution, the data transmission time is negligible.

The waiting times at nodes are independent random variables, each following an exponential distribution $\mathrm{Exp}(\mu)$ with mean $1/\mu$. Thus the average service time

$T_s(\alpha|\varphi(\mathcal{A})) = \frac{1}{\mu}(H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha})$, where $H_\alpha = \sum_{i=1}^{\alpha} 1/i$ is the $\alpha$-th harmonic number [23]. The corresponding service rate achieved by the nodes in $\mathcal{A}$ (with $\varphi(\mathcal{A}) > \alpha$ nodes containing data) is

$$\mu_s(\alpha|\varphi(\mathcal{A})) = \frac{1}{T_s(\alpha|\varphi(\mathcal{A}))} = \frac{\mu}{H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha}}. \quad (2)$$

It is not hard to see that

$$\mu_s(\alpha|\varphi(\mathcal{A})) \leq \mu\varphi(\mathcal{A}). \quad (3)$$

*2) Large File Model:* Here the download time scales with the number of chunks being downloaded. We consider two distribution/scaling models for the service time: scaled exponential and scaled-shift exponential.

Scaled Exponential Service Time: We assume that a node storing the whole file delivers all of its blocks in a random time, exponentially distributed with the mean $1/\mu$, and that a node storing $1/\alpha$ fraction of the file delivers all of its blocks in the random time exponentially distributed with the mean $1/(\alpha\mu)$. For this model, by applying the order statistics for exponential distribution, we have $T_s(\alpha|\varphi(\mathcal{A})) = \frac{1}{\alpha\mu}(H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha})$, where $1/(\alpha\mu)$ comes from the service rate scaling discussed above. The corresponding service rate from set $\mathcal{A}$ is

$$\mu_s(\alpha|\varphi(\mathcal{A})) = \frac{\alpha\mu}{H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha}}. \quad (4)$$

It is not hard to see that

$$\mu\varphi(\mathcal{A}) \geq \mu_s(\alpha|\varphi(\mathcal{A})) \geq \mu(\varphi(\mathcal{A}) - \alpha + 1). \quad (5)$$

Shifted Exponential Service Time: Here the data delivery consists of two steps: first, the node takes an exponential random time to process the request; second, it takes a constant time, proportional to its number of the node's stored data blocks, to deliver them to the user. Therefore, the two-step delivery time for a node storing $1/\alpha$ fraction of the file can be modeled by the shifted exponential distribution with rate $\mu$ and the shift parameter $\Delta/\alpha$, denoted by S-Exp$(\Delta/\alpha, \mu)$. For this model, since shifted exponential distribution is a combination of a constant and an exponential tail, we have $T_s(\alpha|\varphi(\mathcal{A})) = \frac{\Delta}{\alpha} + \frac{1}{\mu}(H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha})$. The corresponding service rate from set $\mathcal{A}$ is

$$\mu_s(\alpha|\varphi(\mathcal{A})) = \frac{\alpha\mu}{\Delta\mu + \alpha(H_{\varphi(\mathcal{A})} - H_{\varphi(\mathcal{A})-\alpha})}. \quad (6)$$

As $\varphi(\mathcal{A}) \in [\alpha, m\alpha]$, it is not hard to see that

$$\frac{\mu\varphi(\mathcal{A})}{\Delta\mu + \alpha} \geq \mu_s(\alpha|\varphi(\mathcal{A})) \geq \frac{\alpha\mu(\varphi(\mathcal{A}) - \alpha + 1)}{\Delta\mu(m\alpha - \alpha + 1) + \alpha^2}. \quad (7)$$

### *D. DSS Performance Metrics*

We consider two key performance metrics: *probability of successful data recovery* and *average service rate*. In general, both measures are of interest and should be simultaneously taken into account. However, as we will see below, they are often maximized by different allocations. In many applications, increasing the chance of successful download is desirable but should not come at the cost of intolerable delivery delay.

*1) Probability of File Recovery:* For an $\alpha$ quasi-uniform allocation, data recovery from this subset is successful iff $\varphi(\mathcal{A}) \geq \alpha$. The probability of successful file recovery under $\alpha$ quasi-uniform allocation is

$$P_s(\alpha) = \sum_{\mathcal{A}:\, \sum_{i\in\mathcal{A}} s_i \geq k} P(\mathcal{A}) \quad (8)$$

where $P(\mathcal{A})$ is the probability of accessing $\mathcal{A}$. Note that the sum goes over all sets $\mathcal{A}$ that satisfy the condition (1). It follows that $P(\mathcal{A}) = \binom{m\alpha}{\varphi(\mathcal{A})}\binom{N-m\alpha}{r-\varphi(\mathcal{A})}/\binom{N}{r}$ for the fixed-size access and $P(\mathcal{A}) = \binom{m\alpha}{\varphi(\mathcal{A})}(1-p)^{\varphi(\mathcal{A})}p^{m\alpha-\varphi(\mathcal{A})}$ for the probabilistic access.

*2) Service Rate:* Under an $\alpha$ quasi-uniform allocation, the service rate $\mu_s(\alpha)$, found by averaging over the conditional service rates, is

$$\mu_s(\alpha) = \sum_{\mathcal{A}:\, \sum_{i\in\mathcal{A}} s_i \geq k} P(\mathcal{A})\mu_s(\alpha|\varphi(\mathcal{A})) \quad (9)$$

where $\mu_s(\alpha|\varphi(\mathcal{A}))$ is the service rate when the set of accessed nodes is $\mathcal{A}$, given by (2), (4) or (6). When the set $\mathcal{A}$ does not satisfy the condition (1), we define $\mu_s(\alpha|\varphi(\mathcal{A})) = 0$.

## III. Problem Statement and Contributions

| | | |
|---|---|---|
| $N$ | – | number of storage nodes (DSS size) |
| $m$ | – | $mk$ is number of encoded blocks |
| $\mathcal{A}$ | – | accessed subset of nodes |
| $\varphi(\mathcal{A})$ | – | number of nodes in $\mathcal{A}$ containing data |
| $\alpha$ | – | $m\alpha$ is the number of nodes with blocks |
| $P_s(\alpha)$ | – | probability of successful file recovery |
| $\mu_s(\alpha)$ | – | DSS service rate |
| $k$ | – | number of block in a file |
| $r$ | – | number of accessed nodes (fixed-size model) |
| $p$ | – | probability of failed access (probabilistic model) |

System parameters and notations are summarised in the above list. Our goal is to characterize the DSS service rate $\mu_s(\alpha)$ for the access and service time models defined in Sec. II-B and II-C. We are in particular interested in finding which $\alpha$ maximizes $\mu_s(\alpha)$. Recall that when $\alpha = 1$, we have the minimal spreading allocation, and when $\alpha = N/m$, we have the maximal spreading allocation. When $1 < \alpha < N/m$, we have an $\alpha$ quasi-uniform allocation.

We conclude that the allocation that maximizes the service rate $\mu_s(\alpha)$ depends on the model. For the small file model, the minimal spreading allocation, i.e. $\alpha = 1$, is always optimal, while for the large file model, this is not the case and it is difficult to determine the optimal allocation. We summarize the regimes where the minimal spreading allocation is optimal and non-optimal for large files in Table I.

The probability of successful recovery, denoted by $P_s(\alpha)$, is another important performance metric [11], [12]. Since $P_s(\alpha)$ and $\mu_s(\alpha)$ may exhibit different trends when varying $\alpha$, we also make comparisons between these two metrics in our numerical analysis to find an overall optimal allocation.

We here put together the relevant results of [1], which focuses on small files, and [2], which focuses on large files.

TABLE I

CONDITIONS FOR THE MINIMAL SPREADING ALLOCATION BEING OPTIMAL/NON-OPTIMAL FOR LARGE FILES

| | | Optimality Conditions | |
|---|---|---|---|
| | | **Scaled Exponential** | **Shifted Exponential** |
| Access | **Fixed-size** | $r \leq \min_{2\leq\alpha\leq r}\{1+\frac{N-1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}\}$ | $r \leq \min_{2\leq\alpha\leq r}\{1+\sqrt[\alpha-1]{\frac{\Delta\mu+\alpha}{\alpha(\Delta\mu m+1)\binom{m\alpha-1}{\alpha-1}}}(N-1)\}$ |
| | **Probabilistic** | $p \geq \max_{\alpha\geq 2}\{1-\frac{1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}\}$ | $p \geq \max_{\alpha\geq 2}\{1-\sqrt[\alpha-1]{\frac{\Delta\mu+\alpha}{\alpha(\Delta\mu m+1)\binom{m\alpha-1}{\alpha-1}}}\}$ |
| | | **Non-optimality Conditions** | |
| Access | **Fixed-size** | $r \geq \min_{2\leq\alpha\leq r}\{\sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}\cdot(N-\alpha+1)+\alpha-1\}$ | $r \geq \min_{2\leq\alpha\leq r}\{\sqrt[\alpha-1]{\frac{\Delta\mu m(m\alpha-\alpha+1)+m\alpha^2}{\alpha(\Delta\mu+1)(m\alpha-\alpha+1)}}\cdot(N-\alpha+1)+\alpha-1\}$ |
| | **Probabilistic** | $p \leq \max_{\alpha\geq 2}\{1-\sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}\}$ | $p \leq \max_{\alpha\geq 2}\{1-\sqrt[\alpha-1]{\frac{m(\Delta\mu(m\alpha-\alpha+1)+\alpha^2)}{\alpha(\Delta\mu+1)(m\alpha-\alpha+1)}}\}$ |

Moreover, we compare the results of these papers and those published in [12], which focuses on the probability of access rather than the service rate as a performance metric. Further contributions include the following:

- We provide a new numerical analysis for each access model for small files to characterize how DSS service rate and recovery probability change with $\alpha$ under different $r$ and $p$.
- We consider several examples for each access model for large files. They are helpful in better understanding the allocation problem. We provide an analysis that results in finding the optimal $\alpha$ in the considered cases.
- We obtain the minimal spreading allocation (non) optimality conditions for large files under each access model.
- We present a new numerical analysis for large files under each access model to show the tradeoffs between DSS service rate and recovery probability.

## IV. STORAGE ALLOCATION FOR SMALL FILES

For the small file model, we assume the service time at each node follows an exponential distribution with the mean $\frac{1}{\mu}$. The service rate for an accessed set of nodes $\mathcal{A}$ is $\mu_s(\alpha|\mathcal{A})$, where $\mu_s(\alpha|\mathcal{A}) > 0$ if $\varphi(\mathcal{A}) \geq \alpha$, and 0 otherwise. Thus, the DSS service rate $\mu_s(\alpha)$ depends on all possible sets $\mathcal{A}$, which satisfy $\sum_{i\in\mathcal{A}} s_i \geq k$. In the following two subsections, we determine the $\mu_s(\alpha)$ for the two considered access models. Some of results in this section were published in [1].

### A. Fixed-Size Access Model

For the fixed-size access model and an exponential service time, the DSS service rate in (9) becomes

$$\mu_s(\alpha) = \frac{\mu}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,\, m\alpha)} \frac{1}{H_\varphi - H_{\varphi-\alpha}} \binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi}. \tag{10}$$

Using (8), the probability of successful file recovery is

$$P_s(\alpha) = \sum_{\varphi=\alpha}^{\min(r,\, m\alpha)} \frac{\binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi}}{\binom{N}{r}}. \tag{11}$$

Here and in the rest of the paper, use $\varphi$ instead of $\varphi(\mathcal{A})$. The following lemma gives the minimal spreading service rate.

*Lemma 1: Under the fixed-size access model with an exponential service time, the service rate of minimal spreading allocation, i.e. $\alpha = 1$, is $\mu_s(1) = \mu m r/N$.*

*Proof:* From (10), we get

$$\mu_s(1) = \frac{\mu}{\binom{N}{r}} \sum_{\varphi=1}^{\min(r,m)} \frac{1}{H_\varphi - H_{\varphi-1}} \binom{m}{\varphi}\binom{N-m}{r-\varphi}.$$

Notice that when $\varphi > m\alpha$, $\binom{m\alpha}{\varphi} = 0$. Thus,

$$\begin{aligned}\mu_s(1) &= \frac{\mu}{\binom{N}{r}} \sum_{\varphi=1}^{r} \varphi \binom{m}{\varphi}\binom{N-m}{r-\varphi} \\ &= \frac{\mu m}{\binom{N}{r}} \sum_{\varphi=1}^{r} \binom{m-1}{\varphi-1}\binom{N-m}{r-\varphi}.\end{aligned}$$

Using Vandermonde's convolution, one can show that $\sum_{\varphi=1}^{r}\binom{m-1}{\varphi-1}\binom{N-m}{r-\varphi} = \binom{N-1}{r-1}$. Therefore, $\mu_s(1) = \frac{\mu m r}{N}$. ∎

We next find an upper bound on $\mu_s(\alpha)$ for any $2 \leq \alpha \leq r$ and compare this bound with $\mu_s(1)$.

*Lemma 2: Under the fixed-size access model and an exponential service time, $\mu_s(\alpha) < \frac{\mu m r}{N}$ for $2 \leq \alpha \leq r$.*

*Proof:* By applying (3) to (10) and using Vandermonde's convolution, we arrive at

$$\begin{aligned}\mu_s(\alpha) &< \frac{\mu}{\alpha\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \varphi \binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi} \\ &< \frac{\mu m}{\binom{N}{r}} \sum_{\varphi=0}^{r-1} \binom{m\alpha-1}{\varphi}\binom{N-m\alpha}{r-1-\varphi} \\ &= \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}} = \frac{\mu m r}{N}.\end{aligned}$$

∎

Lemmas 1 and 2, give the following theorem on the optimality of minimum spreading ($\alpha = 1$).

*Theorem 1: Under the fixed-size access and exponential service, the minimal spreading maximizes the service rate.*

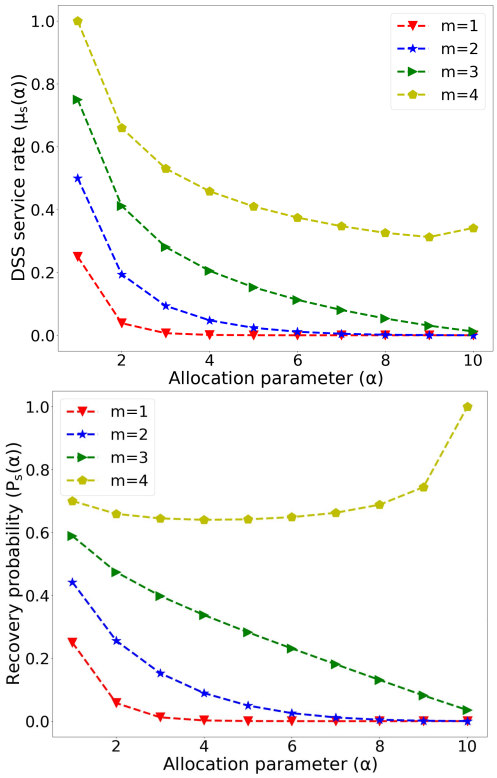



Fig. 2. Comparing the service rate $\mu_s(\alpha)$ (upper, cf.(10)) and the successful recovery probability $P_s(\alpha)$ (lower, cf.(11)) for a range of allocation parameters $\alpha$, under the fixed-size access model. The number of nodes $N$ is 40, and the number of accessed nodes $r$ is 10. The service time follows Exp(1). When $m \leq 3$, the minimal spreading allocation is optimal for both metrics. When $m = 4$, the minimal spreading maximizes the service rate whereas the maximal spreading maximizes the probability of success recovery.

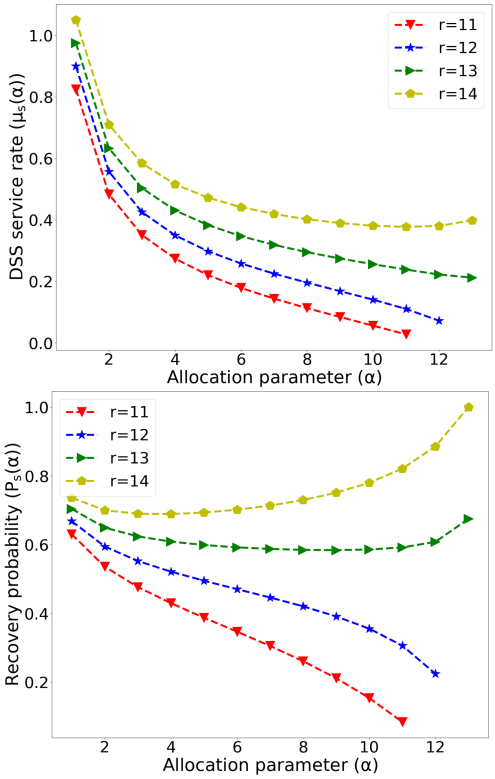



Fig. 3. Comparing the service rate $\mu_s(\alpha)$ (upper, cf. (10)) and the successful recovery probability $P_s(\alpha)$ (lower, cf. (11)) for a range of allocation parameters $\alpha$, under the fixed-size access model. The number of nodes $N$ is 40, and the redundancy level $m$ is 3. The service time follows Exp(1). When $r \leq 13$, the minimal spreading allocation is optimal in both figures. When $r = 14$, $\alpha = 12$ allocation maximizes the successful recovery probability.

*Numerical Analysis:* In Fig. 2, we evaluate (10) and (11) to see how the DSS service rate $\mu_s(\alpha)$ (upper) and the successful recovery probability $P_s(\alpha)$ (lower) changes with the allocation parameter $\alpha$. We consider a system with $N = 40$ storage nodes and $r = 10$ accessed nodes for four different levels of redundancy $m \in \{1, 2, 3, 4\}$. In the upper subfigure, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. When $m \leq 3$, $\mu_s(\alpha)$ decreases with increasing $\alpha$ and approaches 0. When $m = 4$, $\mu_s(\alpha)$ reaches its minimum at $\alpha = 9$. In the lower subfigure, when $m \leq 3$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, thus the minimal spreading allocation is optimal. When $m = 4$, $P_s(\alpha)$ reaches 1 at $\alpha = 10$, i.e. the maximal spreading allocation ($\alpha = N/m$) is optimal. From the observations, we conclude that the minimal spreading allocation is always optimal under the DSS service rate, which is consistent with the result in Theorem 1. The optimal allocation under successful recovery probability is determined by the level of introduced redundancy.

In Fig. 3, we analyze $\mu_s(\alpha)$ vs. $\alpha$ (upper) and $P_s(\alpha)$ (lower) vs. $\alpha$ for different numbers of accessed node $r \in \{10, 11, 12, 13\}$. We consider a system with $N = 40$ storage nodes and a redundancy level of $m = 3$. In the upper subfigure, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. $\mu_s(\alpha)$ decreases with increasing $\alpha$ except for the scenario $r = 14$. In the lower subfigrue, when $r \leq 13$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. When $r = 14$, $P_s(\alpha)$ reaches its maximum at $\alpha = 12$. From the observations, we conclude that the minimal spreading allocation is always optimal under the DSS service rate. However, it is no longer optimal under the successful recovery probability as $r$ increases.

### *B. Probabilistic Access Model*

For probabilistic access model under exponential service time, the DSS service rate (9) becomes

$$\mu_s(\alpha) = \sum_{\varphi=\alpha}^{m\alpha} \frac{\mu}{H_\varphi - H_{\varphi-\alpha}} \binom{m\alpha}{\varphi} (1-p)^\varphi p^{m\alpha-\varphi}. \quad (12)$$

Using (8), the probability of successful file recovery is

$$P_s(\alpha) = \sum_{\varphi=\alpha}^{\min(r,\, m\alpha)} \binom{m\alpha}{\varphi} (1-p)^\varphi p^{m\alpha-\varphi}. \quad (13)$$

We have the following result on the minimal spreading.

*Lemma 3: Under the probabilistic access model with exponential service time, the service rate of the minimal spreading allocation, i.e. $\alpha = 1$, is $\mu_s(1) = \mu m(1-p)$.*

*Proof:* From (12), we get

$$\mu_s(1) = \mu m(1-p) \sum_{\varphi=0}^{m-1} \binom{m-1}{\varphi} (1-p)^\varphi p^{m-\varphi-1}$$

Using binomial expansion, we get $\mu_s(1) = \mu m(1-p)$. ■

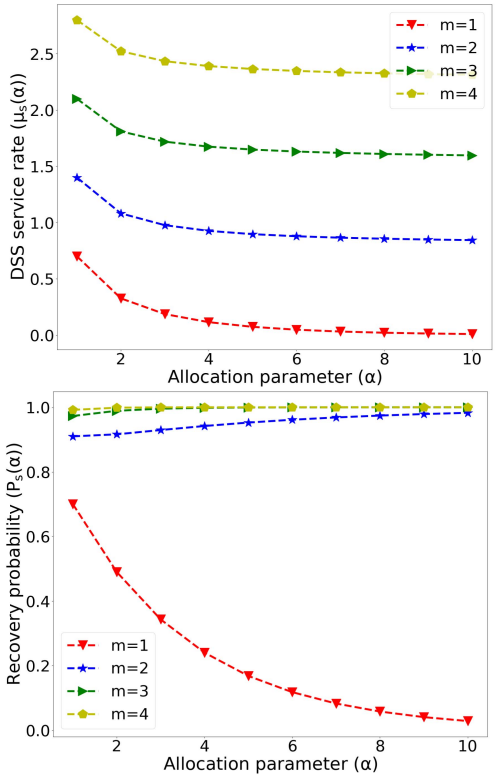



Fig. 4. Comparisons between the service rate $\mu_s(\alpha)$ (upper, cf. (12)) and the successful recovery probability $P_s(\alpha)$ (lower, cf. (13)) as a function of the allocation parameter $\alpha$ under the probabilistic access model. The number of storage nodes is $N = 40$, and the probability of failed access is $p = 0.3$. The service time follows Exp(1). When $m = 1$, the minimal spreading allocation is optimal in both figure. When $m \geq 2$, the minimal spreading allocation is optimal considering $\mu_s(\alpha)$, and performs the worst considering $P_s(\alpha)$.

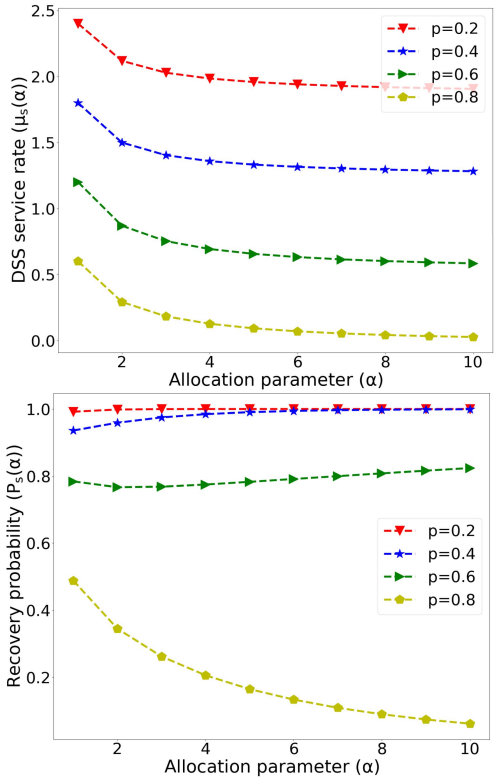



Fig. 5. Comparisons between the DSS service rate $\mu_s(\alpha)$ (upper, cf. (12)) and the successful recovery probability $P_s(\alpha)$ (lower, cf. (13)) as a function of the allocation parameter $\alpha$ under the probabilistic access model. The number of storage nodes is $N = 40$, and the redundancy level is $m = 3$. The service time follows Exp(1). minimal spreading allocation is optimal with regards to $\mu_s(\alpha)$. For $p = 0.8$, the minimal spreading allocation is also optimal for $P_s(\alpha)$ while it performs the worst when $p \leq 0.4$.

Similar to Lemma 2, we find an upper bound on the DSS service rate when $\alpha \geq 2$.

*Lemma 4: Under the probabilistic access model with an exponential service time, for any $\alpha$ quasi-uniform allocation, its service rate satisfies $\mu_s(\alpha) < \mu m(1-p)$.*

*Proof:* By applying (3) to (12), we arrive at

$$\mu_s(\alpha) < \mu m(1-p) \sum_{\varphi=\alpha-1}^{m\alpha-1} \binom{m\alpha-1}{\varphi} (1-p)^{\varphi} p^{m\alpha-\varphi-1}$$

$$< \mu m(1-p) \sum_{\varphi=0}^{m\alpha-1} \binom{m\alpha-1}{\varphi} (1-p)^{\varphi} p^{m\alpha-\varphi-1} = \mu m(1-p).$$

■

Using Lemmas 3 and 4, we have the following result on the optimal storage allocation for the probabilistic access model.

*Theorem 2: Under the probabilistic access model with an exponential service time, the minimal spreading allocation maximizes the DSS service rate.*

*Numerical Analysis:* In Fig. 4, we respectively evaluate (12) and (13) to see how the DSS service rate $\mu_s(\alpha)$ (upper) and the successful recovery probability $P_s(\alpha)$ (lower) changes with the allocation parameter $\alpha$. We consider a system with $N = 40$ storage nodes and a failed access probability $p = 0.3$ for four different levels of redundancy $m \in \{1, 2, 3, 4\}$. In the upper subfigure, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal, and decreases with increasing $\alpha$. In the lower subfigure, when $m = 1$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. When $m \geq 2$, $P_s(\alpha)$ reaches its maximum at $\alpha = 10$, approaching 1 for $m = 3$ and 4 when $\alpha \geq 4$. From the observations, we conclude that the minimal spreading allocation is always optimal under the DSS service rate, which is consistent with the result in Theorem 2. The optimal allocation under successful recovery probability is determined by the level of introduced redundancy.

In Fig. 5, we analyze $\mu_s(\alpha)$ vs. $\alpha$ (upper) and $P_s(\alpha)$ (lower) vs. $\alpha$ for $p \in \{0.2, 0.4, 0.6, 0.8\}$. We consider a system with $N = 40$ storage nodes and a redundancy level of $m = 3$. In the upper subfigure, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. $\mu_s(\alpha)$ decreases with increasing $\alpha$. In the lower subfigrue, when $p \leq 0.4$, $P_s(\alpha)$ increases with $\alpha$, and reaches 1 at about $\alpha \geq 6$. When $p = 0.6$, $P_s(\alpha)$ takes values around 0.8. When $p = 0.8$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$. From the observations in Fig. 5, we conclude that the minimal spreading allocation is always optimal under the DSS service rate. However, the optimal allocation under the successful recovery probability depends on the probability of failed access.

## V. Storage Allocation for Large Files With Scaled Exponential Service Time

For the large file model, we assume that the service time at each node follows a scaled exponential distribution with the mean $\frac{1}{\alpha\mu}$, i.e. the allocation parameter $\alpha$ changes the scale of an exponential distribution. In the following two subsections, we determine the $\mu_s(\alpha)$ for the two considered access models. Some of the results in this section were published in [2].

### A. Fixed-Size Access Model

For the fixed-size access model under a scaled exponential service time, the DSS service rate (9) becomes

$$\mu_s(\alpha) = \frac{\mu\alpha}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \frac{1}{H_\varphi - H_{\varphi-\alpha}} \binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi}. \tag{14}$$

By comparing (10) and (14), we expect the minimal spreading to not always be optimal. We prove that the maximal spreading performs better under some system parameters' values.

*Lemma 5: Under fixed-size access model with a scaled exponential service time, the service rate of the maximal spreading allocation, i.e. $\alpha = r$, is $\mu_s(r) = \frac{\mu r\binom{rm}{r}}{H_r\binom{N}{r}}$. The service rate of the minimal spreading allocation, i.e. $\alpha = 1$, is $\mu_s(1) = \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}} = \frac{\mu m r}{N}$.*

*Proof:* From (14), for the maximal spreading allocation, we get $\mu_s(r) = \frac{\mu r}{\binom{N}{r}} \sum_{\varphi=r}^{\min(r,rm)} \frac{1}{H_\varphi - H_{\varphi-r}} \binom{rm}{\varphi}\cdot\binom{\text{N-rm}}{r-\varphi}$. Since $m \geq 1$ and $H_0 = 0$, $\mu_s(r) = \frac{\mu r}{\binom{N}{r}}\frac{1}{H_r}\binom{rm}{r}\binom{\text{N-rm}}{r-r} = \frac{\mu r\binom{rm}{r}}{H_r\binom{N}{r}}$.

For the minimal spreading allocation, we get $\mu_s(1|\varphi) = \mu\varphi$. Thus, by applying the same approach as in the proof of Lemma 1, we get $\mu_s(1) = \frac{\mu m r}{N} = \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}}$. ∎

With Lemma 5, we obtain the following result.

*Theorem 3 (Minimal vs. Maximal Spreading): Under the fixed-size access model with a scaled exponential service time, when $rm \geq N$, the maximal spreading allocation outperforms the minimal spreading allocation, i.e. $\mu_s(r) \geq \mu_s(1)$.*

*Proof:* Lemma 5 with $rm \geq N$ and $r \geq H_r$ gives $\mu_s(r) = \frac{\mu r\binom{rm}{r}}{H_r\binom{N}{r}} = \frac{\mu r m\binom{rm-1}{r-1}}{H_r\binom{N}{r}} \geq \frac{\mu m H_r\binom{N-1}{r-1}}{H_r\binom{N}{r}} = \frac{\mu m r}{N} = \mu_s(1)$. ∎

*Remark 1: Under the fixed-size access model with a scaled exponential service time, the minimal spreading allocation does not always maximize the DSS service rate.*

*1) Optimal and Non-Optimality Conditions for the Minimal Spreading Allocation:* Considering the complexity of (14), finding the $\alpha$ that maximizes $\mu_s(\alpha)$ is hard. Instead, we find the optimality and non-optimality conditions for the minimal spreading allocation in the following.

*Theorem 4 (Minimal Spreading Optimality Condition): Under the fixed-size access with a scaled exponential service, the minimal spreading maximizes the service rate $\mu_s(\alpha)$ when $r \leq \min_{2\leq\alpha\leq r}\{1 + \frac{N-1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}\}$.*

*Proof:* We need to find the condition ensuring $\mu_s(1) \geq \mu_s(\alpha)$ for all $2 \leq \alpha \leq r$. Consider $\mu_s(\alpha)$ in (14). By (5), we have $\mu_s(\alpha|\varphi) < \mu\varphi$ when $\alpha \geq 2$, and thus

$$\begin{aligned}\mu_s(\alpha) &< \frac{\mu}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \varphi\binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi} \\ &= \frac{\mu m\alpha}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \prod_{i=0}^{\alpha-2} \frac{m\alpha-1-i}{\varphi-1-i}\binom{m\alpha-\alpha}{\varphi-\alpha} \\ &\quad\times\binom{N-m\alpha}{r-\varphi}.\end{aligned}$$

Since $\varphi$ goes from $\alpha$ to $m\alpha$, we further have

$$\begin{aligned}&\mu_s(\alpha) \\ &< \frac{\mu m\alpha}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \Big(\prod_{i=0}^{\alpha-2} \frac{m\alpha-1-i}{\alpha-1-i}\Big)\binom{m\alpha-\alpha}{\varphi-\alpha} \\ &\quad\times\binom{N-m\alpha}{r-\varphi} \\ &= \frac{\mu m\alpha\binom{m\alpha-1}{\alpha-1}\binom{N-\alpha}{r-\alpha}}{\binom{N}{r}} \text{ (by Vandermonde's convolution).}\end{aligned}$$

According to Lemma 3, we have $\mu_s(1) = \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}}$. To satisfy $\mu_s(\alpha) \leq \mu_s(1)$, we have

$$\begin{aligned}\frac{\mu m\alpha\binom{m\alpha-1}{\alpha-1}\binom{N-\alpha}{r-\alpha}}{\binom{N}{r}} \leq \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}} &\Leftrightarrow \alpha\binom{m\alpha-1}{\alpha-1} \\ &\leq \prod_{i=1}^{\alpha-1}\frac{N-i}{r-i}.\end{aligned} \tag{15}$$

As $\frac{N-i}{r-i} < \frac{N-1-i}{r-1-i}$ for $N > r$, it can be shown that $\prod_{i=1}^{\alpha-1}\frac{N-i}{r-i} > (\frac{N-1}{r-1})^{\alpha-1}$, and as a result, (15) holds when

$$\alpha\binom{m\alpha-1}{\alpha-1} \leq (\frac{N-1}{r-1})^{\alpha-1} \Leftrightarrow r \leq 1 + \frac{N-1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}. \tag{16}$$

If (16) holds for all $2 \leq \alpha \leq r$, $\mu_s(1)$ is optimal. ∎

*Theorem 5 (Minimal Spreading Non-Optimality Condition): Under the fixed-size access model with a scaled exponential service time, the minimal spreading allocation does not maximize the DSS service rate $\mu_s(\alpha)$ when $r \geq \min_{2\leq\alpha\leq r}\{\sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}(N-\alpha+1)+\alpha-1\}$.*

*Proof:* To prove the theorem statement, we need to find the condition ensuring $\mu_s(1) \leq \mu_s(\alpha)$ for at least one $\alpha \in [2, r]$. The expression of $\mu_s(\alpha)$ is given in (14). According to (5), we have $\mu_s(\alpha|\varphi) > \mu(\varphi-\alpha+1)$ when $\alpha \geq 2$. Thus,

$$\begin{aligned}\mu_s(\alpha) &> \frac{\mu}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} (\varphi-\alpha+1)\binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi} \\ &= \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \frac{\mu(\varphi-\alpha+1)}{\binom{N}{r}} \prod_{i=0}^{\alpha-2}\frac{m\alpha-i}{\varphi-i} \\ &\quad\times\binom{m\alpha-\alpha+1}{\varphi-\alpha+1}\binom{N-m\alpha}{r-\varphi}.\end{aligned}$$

Since $\varphi$ goes from $\alpha$ to $m\alpha$, we further have

$$\mu_s(\alpha) > \frac{\mu}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} (\varphi-\alpha+1)\binom{m\alpha-\alpha+1}{\varphi-\alpha+1}\binom{N-m\alpha}{r-\varphi} = \frac{\mu(m\alpha-\alpha+1)\binom{N-\alpha}{r-\alpha}}{\binom{N}{r}} \text{ (Vandermonde's convolution).}$$

According to Lemma 3, we have $\mu_s(1) = \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}}$, hence, to satisfy $\mu_s(\alpha) \geq \mu_s(1)$, we need to have

$$\frac{\mu(m\alpha-\alpha+1)\binom{N-\alpha}{r-\alpha}}{\binom{N}{r}} \geq \frac{\mu m\binom{N-1}{r-1}}{\binom{N}{r}} \Leftrightarrow \frac{m\alpha-\alpha+1}{m} \geq \prod_{i=0}^{\alpha-2}\frac{N-1-i}{r-1-i} \tag{17}$$

Since $\frac{N-1-i}{r-1-i} < \frac{N-2-i}{r-2-i}$ for $N > r$, we have $\prod_{i=0}^{\alpha-2}\frac{N-1-i}{r-1-i} < (\frac{N-\alpha+1}{r-\alpha+1})^{\alpha-1}$. Hence, if $\frac{m\alpha-\alpha+1}{m} \geq (\frac{N-\alpha+1}{r-\alpha+1})^{\alpha-1} \Leftrightarrow r \geq \sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}(N-\alpha+1)+\alpha-1$, the inequality (17) holds, meaning that $\mu_s(\alpha) \geq \mu_s(1)$ and $\mu_s(1)$ is not optimal. ■

From Theorems 4 and 5, we see that both conditions depend on the number of accessed nodes $r$. Roughly speaking, when $r$ is small, $\alpha = 1$ is optimal while an $\alpha > 1$ is optimal when $r$ is large. Based on these theorems, we conjecture the following about the value of $\alpha$ that maximizes $\mu_s(\alpha)$.

*Conjecture 1: The following assertions are true*:

1) *Under the fixed-size access model, given the number of nodes $N$ and the redundancy level $m$, the optimal $\alpha$ increases as $r$ increases.*
2) *For every $N$ and $m$, there exists a $\gamma \in (1, N)$, such that for all $r \leq \gamma$, the minimal spreading is optimal.*
3) *For every $N$ and $m$, there exists a $\zeta \in (1, N)$, such that for all $r \geq \zeta$, the maximal spreading is optimal.*

*2) Numerical Analysis:* In Fig. 6, we evaluate the expression of $\mu_s(\alpha)$ given in (14) to see how the DSS service rate changes with the allocation parameter $\alpha$. We consider a system with $N = 40$ storage nodes. Using Theorems 4 and 5, we can easily calculate the optimality and non-optimality conditions. For example, when $m = 2$, the minimal spreading allocation is optimal when $r \leq 7.5$ and is non-optimal when $r > 27$. These conditions provide some knowledge of the optimal allocation and further insight on the optimal allocation can be found from Fig. 6. The upper graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different level of redundancy $m \in \{1, 2, 3, 4\}$, and the number of accessed nodes $r = 10$. The lower graph shows $\mu_s(\alpha)$ vs. $\alpha$ for different numbers of accessed nodes $r \in \{8, 10, 12, 13\}$, and the redundancy level $m = 3$. In the upper subfigure, when $m \leq 2$, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 1$ and decreases with increasing $\alpha$, i.e. the minimal spreading allocation is optimal. When $m = 3$, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 3$. When $m = 4$, $\mu_s(\alpha)$ increases with $\alpha$ and reaches its maximum at $\alpha = 10$, i.e. the maximal spreading allocation ($\alpha = N/m$) is optimal. In the lower subfigure, when $r = 8$, the minimal spreading allocation is optimal, while this is not the case for $r \geq 10$. Since the redundancy level is $m = 3$, which means $N/m$ is not an integer, we cannot apply the maximal spreading allocation. When $r = 13$, the optimal allocation is at $\alpha = 13$, where we allocate the file into 39 of 40 nodes. From the observations in Fig. 6, we conclude that the minimal spreading allocation is optimal only when $m$ or $r$ is sufficiently small. By increasing either of the parameters, an $\alpha \geq 2$ quasi-uniform allocation becomes optimal, and when $m$ or $r$ is sufficiently large, the maximal spreading allocation is optimal.

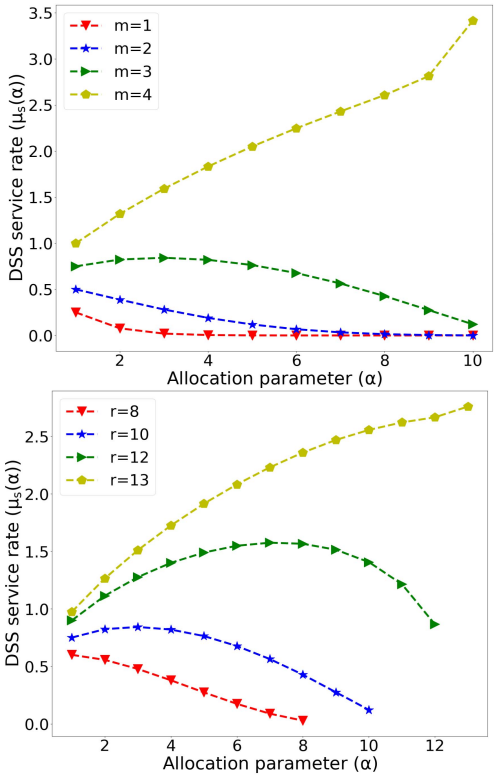



Fig. 6. The DSS service rate $\mu_s(\alpha)$ for the fixed-size access model with scaled exponential service time as a function of the allocation parameter $\alpha$ (cf. (14)). The number of storage nodes is $N = 40$, and the service time follows $\text{Exp}(1/\alpha)$. (upper) $\mu_s(\alpha)$ vs. $\alpha$ with $r = 10$ accessed nodes for four values of $m$. (lower) $\mu_s(\alpha)$ vs. $\alpha$ with $m = 3$ redundancy for four values of $r$. Given $r$ (or $m$), the optimal allocation changes from the minimal spreading allocation to the maximal spreading allocation as $m$ (or $r$) increases.

In Fig. 7, we analyze $P_s(\alpha)$ vs. $\mu_s(\alpha)$ as $\alpha$ increases from 1 to $r$. We consider a system with $N = 40$ storage nodes and two values for each parameter, i.e. $m \in \{3, 4\}$ and $r \in \{8, 10\}$. Some observations can be made from the figure: when both $m$ and $r$ are sufficiently small, the minimal spreading allocation is optimal for both $P_s(\alpha)$ and $\mu_s(\alpha)$. When both $m$ and $r$ are sufficiently large, the maximal spreading allocation is optimal for both performance metrics. Otherwise, we cannot find an optimal allocation to simultaneously optimize both performance metrics. For example, when $m = 4$ and $r = 8$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, while $\mu_s(\alpha)$ reaches its maximum at $\alpha = 4$. A performance tradeoff can be achieved by choosing $\alpha = 2$ to obtain acceptable $P_s(\alpha)$ and $\mu_s(\alpha)$.

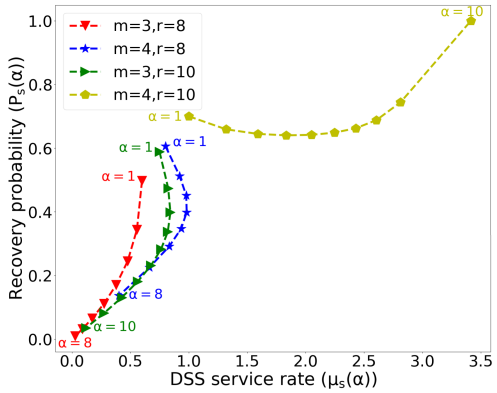



Fig. 7. Successful recovery probability $P_s(\alpha)$ vs. the DSS service rate $\mu_s(\alpha)$ for the fixed-size access model as a function of $\alpha$ for different values of $m$ and $r$ (cf. (11) and (14)). The number of storage nodes is $N = 40$, and the service time follows $\mathrm{Exp}(1/\alpha)$. The minimal (or maximal) spreading allocation is optimal when $m$ and $r$ are sufficiently small (or large). Otherwise, the optimal allocation is different for different performance metrics.

### *B. Probabilistic Access Model*

For probabilistic access model under scaled exponential service time, the DSS service rate (9) becomes

$$\mu_s(\alpha) = \sum_{\varphi=\alpha}^{m\alpha} \frac{\mu\alpha}{H_\varphi - H_{\varphi-\alpha}} \binom{m\alpha}{\varphi} (1-p)^\varphi p^{m\alpha-\varphi}. \quad (18)$$

By comparing (12) and (18), we expect that the minimal spreading allocation may not be always optimal. To this end, we first present the following result on the optimal $\alpha$ to maximize the service rate when no redundancy is used.

*Theorem 6 (Optimal $\alpha$): Under the probabilistic access model with a scaled exponential service time and considering a no redundancy scenario, i.e. $m = 1$, the optimal $\alpha$ which maximizes the DSS service rate $\mu_s(\alpha)$ is located in the range $[(1/2-p)/p, (1-p)/p]$.*

*Proof:* Since the value of $\alpha$ is an integer, $\mu_s(\alpha)$ as a function of $\alpha$ is discrete. To prove the optimal $\alpha$ is located in the range $[(1/2-p)/p, (1-p)/p]$, we need to show $\mu_s(\alpha)$ increases with $\alpha$ when $\alpha \le (1/2-p)/p$ (i.e. $\mu_s(\alpha) \le \mu_s(\alpha+1)$) and decreases with increasing $\alpha$ when $\alpha \ge (1-p)/p$ (i.e. $\mu_s(\alpha) \ge \mu_s(\alpha+1)$). For $\alpha \ge 1$, the DSS service rate is $\mu_s(\alpha) = \frac{\mu\alpha}{H_\alpha}(1-p)^\alpha$, resulting in $\mu_s(\alpha)/\mu_s(\alpha+1) = (\alpha H_{\alpha+1})/((\alpha+1)H_\alpha(1-p))$. Thus,

$$\frac{\mu_s(\alpha)}{\mu_s(\alpha+1)} \ge 1 \Leftrightarrow \frac{\alpha H_{\alpha+1}}{(\alpha+1)H_\alpha(1-p)} \ge 1$$
$$\Leftrightarrow \frac{\alpha}{\alpha+1} \ge (1-\alpha p-p)H_\alpha. \quad (19)$$

Since $\frac{\alpha}{\alpha+1} > 0$, (19) is satisfied when $(1-\alpha p-p)H_\alpha \le 0$, resulting in $\alpha \ge (1-p)/p$. Similarly,

$$\frac{\mu_s(\alpha)}{\mu_s(\alpha+1)} \le 1 \Leftrightarrow \frac{\alpha}{\alpha+1} \le (1-\alpha p-p)H_\alpha. \quad (20)$$

On the other hand, since $\frac{1}{2}H_\alpha - \frac{\alpha}{\alpha+1} = \frac{1}{2}(H_{\alpha+1} + \frac{1}{\alpha+1}) - 1 = \frac{1}{2}(\sum_{i=2}^{\alpha+1}\frac{1}{i} + \frac{1}{\alpha+1}) - \frac{1}{2} \ge \frac{1}{2}(\frac{\alpha}{\alpha+1} + \frac{1}{\alpha+1}) - \frac{1}{2} = 0$, we have $\frac{\alpha}{\alpha+1} \le 1/2H_\alpha$. Thus, if $(1-\alpha p-p) \ge 1/2$, or equivalently $\alpha \le (1/2-p)/p$, (20) holds. ■

From Theorem 6, it is easy to see that $p$ is small, the optimal $\alpha$ is greater than 1. Then we have the following remark on the minimal spreading allocation.

*Remark 2: Under the probabilistic access model with a scaled exponential service time, the minimal spreading allocation ($\alpha = 1$) does not always maximize the service rate.*

*1) Optimality and Non-Optimality Conditions for Minimal Spreading Allocation:* From Remark 2, we know that the minimal spreading allocation is not always optimal. Considering the complexity of (18), finding the $\alpha$ that maximizes $\mu_s(\alpha)$ is hard. Similar to the fixed-size access model in Sec. V-A.1, we find the optimality and non-optimality conditions for the minimal spreading allocation.

*Theorem 7 (Minimal Spreading Optimality Condition): Under the probabilistic access model with a scaled exponential service time, the minimal spreading allocation maximizes the DSS service rate $\mu_s(\alpha)$ when $p \ge \max_{\alpha\ge2}\{1 - \frac{1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}\}$.*

*Proof:* To prove the theorem statement, we need to find the condition ensuring $\mu_s(1) \ge \mu_s(\alpha)$ for all $\alpha \ge 2$. Using (18) and considering that $\mu_s(\alpha|\varphi) < \mu\varphi$ for $\alpha \ge 2$ according to (5), we have

$$\begin{aligned}
&\mu_s(\alpha) \\
&< \mu \sum_{\varphi=\alpha}^{m\alpha} \varphi \binom{m\alpha}{\varphi} (1-p)^\varphi p^{m\alpha-\varphi} \\
&= \mu m\alpha \sum_{\varphi=\alpha}^{m\alpha} \Big(\prod_{i=0}^{\alpha-2} \frac{m\alpha-1-i}{\varphi-1-i}\Big) \binom{m\alpha-\alpha}{\varphi-\alpha} (1-p)^\varphi p^{m\alpha-\varphi}.
\end{aligned}$$

Since $\varphi$ goes from $\alpha$ to $m\alpha$, we have

$$\begin{aligned}
&\mu_s(\alpha) \\
&< \mu m\alpha \binom{m\alpha-1}{\alpha-1} (1-p)^\alpha \cdot \\
&\quad \sum_{\varphi=0}^{m\alpha-\alpha} \binom{m\alpha-\alpha}{\varphi} (1-p)^\varphi p^{m\alpha-\alpha-\varphi} \\
&= \mu m\alpha \binom{m\alpha-1}{\alpha-1} (1-p)^\alpha \quad \text{(by binomial expansion)}.
\end{aligned}$$

From (18), $\mu_s(1) = \mu\sum_{\varphi=1}^{m}\varphi\binom{m}{\varphi}(1-p)^\varphi p^{m-\varphi} = \mu m(1-p)$. Now, to satisfy $\mu_s(\alpha) \le \mu_s(1)$, we have $\mu m\alpha\binom{m\alpha-1}{\alpha-1}(1-p)^\alpha \le \mu m(1-p) \Leftrightarrow (1-p)^{\alpha-1} \le \frac{1}{\alpha\binom{m\alpha-1}{\alpha-1}}$. Thus, if $p \ge 1 - \frac{1}{\sqrt[\alpha-1]{\alpha\binom{m\alpha-1}{\alpha-1}}}$ holds for all $\alpha \ge 2$, $\mu_s(1)$ is optimal. ■

*Theorem 8 (Minimal Spreading Non-Optimality Condition): Under the probabilistic access model with a scaled exponential service time, the minimal spreading allocation does not maximize the DSS service rate $\mu_s(\alpha)$ when $p \le \max_{\alpha\ge2}\{1 - \sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}\}$.*

*Proof:* We are interested in finding a condition that guarantees the existence of an $\alpha \ge 2$, such that $\mu_s(1) \le \mu_s(\alpha)$. The expression for $\mu_s(\alpha)$ is given in (18). According

to (5), we have $\mu_s(\alpha|\varphi) > \mu(\varphi-\alpha+1)$ when $\alpha \geq 2$, then

$$\begin{aligned}\mu_s(\alpha) &> \mu \sum_{\varphi=\alpha}^{m\alpha} (\varphi-\alpha+1)\binom{m\alpha}{\varphi}(1-p)^{\varphi}p^{m\alpha-\varphi} \\ &= \mu \sum_{\varphi=\alpha}^{m\alpha} (\varphi-\alpha+1)\binom{m\alpha-\alpha+1}{\varphi-\alpha+1} \\ &\quad \times (1-p)^{\varphi}p^{m\alpha-\varphi} \\ &\quad \times \prod_{i=0}^{\alpha-2} \frac{m\alpha-i}{\varphi-i}\end{aligned}$$

Since $\varphi$ goes from $\alpha$ to $m\alpha$, thus

$$\begin{aligned}\mu_s(\alpha) &> \mu \sum_{\varphi=\alpha}^{m\alpha} (\varphi-\alpha+1)\binom{m\alpha-\alpha+1}{\varphi-\alpha+1}(1-p)^{\varphi}p^{m\alpha-\varphi} \\ &= \mu(m\alpha-\alpha+1)(1-p)^{\alpha} \sum_{\varphi=0}^{m\alpha-\alpha} \binom{m\alpha-\alpha}{\varphi} \\ &\quad \times (1-p)^{\varphi}p^{m\alpha-\alpha-\varphi} \\ &= \mu(m\alpha-\alpha+1)(1-p)^{\alpha}.\end{aligned}$$

Since $\mu_s(1) = \mu m(1-p)$, to satisfy $\mu_s(\alpha) \geq \mu_s(1)$, it is sufficient to have $\mu(m\alpha-\alpha+1)(1-p)^{\alpha} \geq \mu m(1-p) \Leftrightarrow (1-p)^{\alpha-1} \geq \frac{m}{m\alpha-\alpha+1}$. Thus, if $p \leq 1 - \sqrt[\alpha-1]{\frac{m}{m\alpha-\alpha+1}}$ holds for an $\alpha \geq 2$, $\mu_s(1)$ is not optimal. ■

From Theorems 7 and 8, we see that both the optimality and non-optimality conditions for the minimal spreading allocation depend on the probability of failed access $p$. Roughly speaking, when $p$ is close to 1, then $\alpha = 1$ is optimal while for $p$ close to 0, an $\alpha > 1$ is optimal. Based on these theorems, we conjecture the following about $\alpha$ that maximizes $\mu_s(\alpha)$.

*Conjecture 2: The following assertions are true*:

1) *Under the probabilistic access model, given the redundancy level $m$, the optimal $\alpha$ increases as $p$ decreases.*
2) *For every $m$, there exists a $\gamma \in (0,1)$, such that for all $p \geq \gamma$, the minimal spreading is optimal.*
3) *For every $m$, there exists a $\zeta \in (0,1)$, such that for all $p \leq \zeta$, the maximal spreading is optimal.*

*Remark 3: The derived bounds on $p$ that guarantee the optimality or non-optimality of minimal spreading are not tight. Thus, there exists a gap between these bounds. Consider, for example, the probabilistic access model and $m = 2$. By applying the conditions in Theorem 7 and 8, we arrive at an optimality sufficient condition of $p \geq 0.83$ and a non-optimality sufficient condition of $p \leq 0.33$.*

*2) Numerical Analysis:* In Fig. 8, we evaluate the expression in (18) for $\mu_s(\alpha)$ to see how the DSS service rate changes with $\alpha$. We consider a system with $N \geq m\alpha$ storage nodes. Using Theorems 7 and 8, we can easily calculate the optimality and non-optimality conditions. For example, when $m = 2$, the minimal spreading allocation is optimal when $p \geq 0.83$ and is non-optimal when $p \leq 0.3$. These conditions provide some knowledge of the optimal allocation and further insight on the optimal allocation can be found from Fig. 8. The upper graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different levels of redundancy $m \in \{1,2,3,4\}$, and the failed access probability $p = 0.3$.

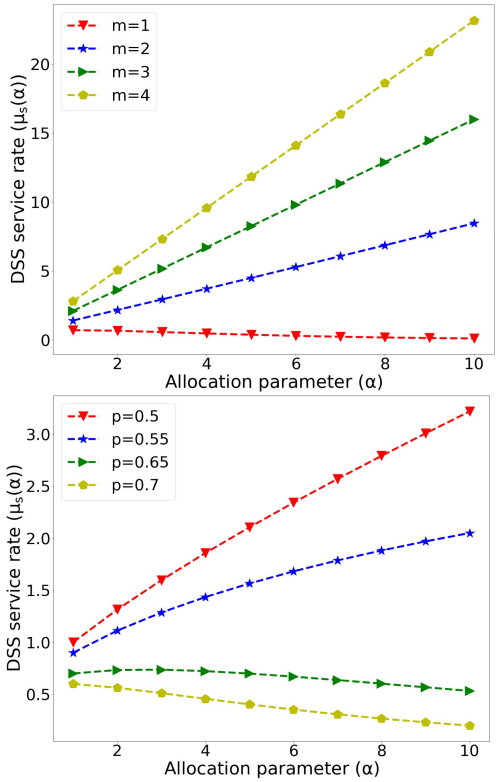



Fig. 8. The DSS service rate $\mu_s(\alpha)$ for the probabilistic access model with scaled exponential service time as a function of the allocation parameter $\alpha$ (cf. (18)). The number of storage nodes is $N \geq m\alpha$, and the service time follows $\text{Exp}(1/\alpha)$. (upper) $\mu_s(\alpha)$ vs. $\alpha$ with the failed access probability $p = 0.3$ for four values of $m$. (lower) $\mu_s(\alpha)$ vs. $\alpha$ with redundancy of $m = 2$ for four values of $p$. Given $p$ (or $m$), the optimal allocation changes from the minimal spreading allocation to the maximal spreading allocation as $m$ increases or $p$ decreases.

The lower graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different values of $p \in \{0.5, 0.55, 0.65, 0.7\}$, and the redundancy level $m = 2$. In the upper subfigure, when $m = 1$, $\mu_s(\alpha)$ decreases with increasing $\alpha$ and reaches its maximum at $\alpha = 1$, i.e. the minimal spreading allocation is optimal. When $m \geq 2$, $\mu_s(\alpha)$ increases with $\alpha$ and reaches its maximum at $\alpha = 10$, i.e. the maximal spreading allocation is optimal. In the lower subfigure, when $p \leq 0.55$, the maximal spreading allocation is optimal. When $p = 0.65$, $\alpha = 2$ allocation is optimal. When $p = 0.7$, the minimal spreading allocation is optimal. Therefore, the optimal allocation changes with $p$.

In Fig. 9, we analyze $P_s(\alpha)$ vs. $\mu_s(\alpha)$ as $\alpha$ increases from 1 to 10. We consider a system with $N \geq m\alpha$ storage nodes and two values for each parameter $m \in \{2,3\}$ and $p \in \{0.45, 0.7\}$. When $m$ is sufficiently small and $p$ is sufficiently large, the minimal spreading allocation is optimal for both $P_s(\alpha)$ and $\mu_s(\alpha)$, while for sufficiently large $m$ and sufficiently small $p$, the maximal spreading allocation is optimal for both performance metrics. For a general scenario, the optimal $\alpha$ for maximizing the service rate may be different from that for maximizing the probability of recovery. For example, when $m = 3$ and $p = 0.7$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, and $\mu_s(\alpha)$ reaches its maximum at $\alpha = 10$.

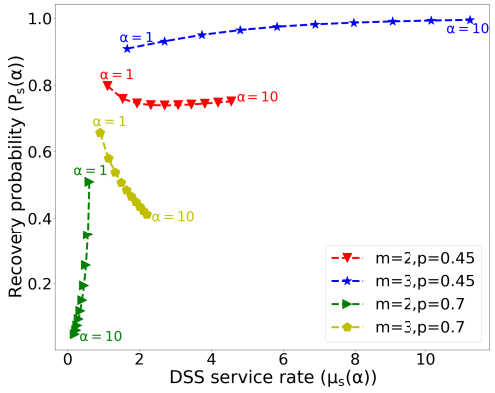



Fig. 9. The successful recovery probability $P_s(\alpha)$ vs. the DSS service rate $\mu_s(\alpha)$ for the probabilistic access model as a function of $\alpha$ for different values of $m$ and $r$ (cf. (13) and (18)). The number of storage nodes is $N \geq m\alpha$, and the service time follows $\text{Exp}(1/\alpha)$. The optimal allocation is affected by both $m$ and $r$ values.

## VI. Storage Allocation for Large Files With Shifted Exponential Service Time

Here the service time follows a shifted exponential distribution with the shift $\Delta/\alpha$ and the rate $\mu$, i.e., S-Exp$(\Delta/\alpha, \mu)$. We determine the $\mu_s(\alpha)$ for the two considered access models. Some of the results in this section were published in [2].

### A. Fixed-Size Access Model

For fixed-size access model under shifted exponential service time, the DSS service rate (9) becomes

$$\mu_s(\alpha) = \frac{\mu\alpha}{\binom{N}{r}} \sum_{\varphi=\alpha}^{\min(r,m\alpha)} \frac{\binom{m\alpha}{\varphi}\binom{N-m\alpha}{r-\varphi}}{\Delta\mu + \alpha(H_\varphi - H_{\varphi-\alpha})} \tag{21}$$

According to (21), the minimal spreading allocation may not be always optimal. In fact, we find a special scenario where there exists an optimal $\alpha \geq 2$ that maximizes the service rate.

Let us start by assuming a constant service time of $\Delta$ for $k$ block. In this case, $\mu_s(\alpha|\varphi(\mathcal{A})) = \alpha/\Delta$ for a set $\mathcal{A}$. Therefore, for the maximal spreading allocation, we have $\mu_s(r) = \frac{r\binom{rm}{r}}{\Delta\binom{N}{r}}$. For the minimal spreading allocation, i.e. $\alpha = 1$, using Vandermonde's convolution, we arrive at $\mu_s(1) = \frac{1}{\Delta}(1 - \frac{\binom{N-m}{r}}{\binom{N}{r}})$.

*Theorem 9 (Minimal vs. Maximal Spreading): Under the fixed-size access model and a constant service time $\Delta$, the maximal spreading allocation outperforms the minimal spreading allocation when $rm \geq N$.*

*Proof:* Since $rm \geq N$ and $\binom{N-m}{r} \geq 0$, $\mu_s(r) = \frac{r\binom{rm}{r}}{\Delta\binom{N}{r}} \geq \frac{\binom{rm}{r} - \binom{N-m}{r}}{\Delta\binom{N}{r}} \geq \frac{\binom{N}{r} - \binom{N-m}{r}}{\Delta\binom{N}{r}} = \mu_s(1)$. ∎

*Remark 4: Under the fixed-size access model with a constant service time $\Delta$, the minimal spreading allocation ($\alpha = 1$) does not always maximize the service rate.*

The constant service time is a special case of the shifted exponential service time. We next find the optimality and non-optimality conditions for the minimal spreading allocation under shifted exponential service time. Our findings show that Remark 4 is also true for the shifted exponential service time.

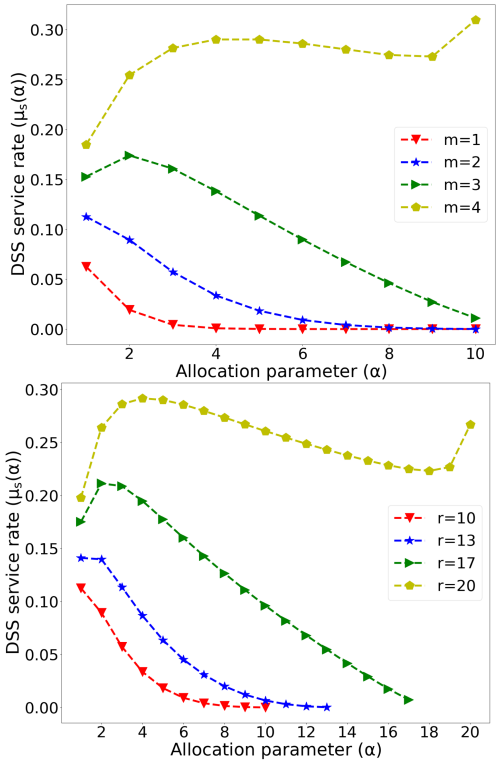



Fig. 10. Service rate $\mu_s(\alpha)$ for the fixed-size access model with shifted exponential service time as a function of the allocation parameter $\alpha$ (cf. (21)). The number of storage nodes is $N = 40$, and the service time follows S-Exp$(3, 1)$. (upper) $\mu_s(\alpha)$ vs. $\alpha$ with $r = 10$ accessed nodes for four values of $m$. (lower) $\mu_s(\alpha)$ vs. $\alpha$ with $m = 2$ redundancy for four values of $r$. Given $r$ (or $m$), the optimal allocation changes from the minimal spreading allocation to the maximal spreading allocation as $m$ (or $r$) increases.

*1) Optimality and Non-Optimality Conditions for the Minimal Spreading Allocation:* finding the optimal $\alpha$ that maximizes $\mu_s(\alpha)$ in (21) is difficult. The proofs of the following theorems are similar to the proofs of Theorem 4 and Theorem 5; see the supplementary Appendix.

*Theorem 10 (Minimal Spreading Optimality Condition): Under the fixed-size access model a shifted exponential service time, the minimal spreading allocation maximizes the DSS service rate $\mu_s(\alpha)$ when*

$$r \leq \min_{2\leq\alpha\leq r} \{1 + \sqrt[\alpha-1]{\frac{\Delta\mu + \alpha}{\alpha(\Delta\mu m + 1)\binom{m\alpha-1}{\alpha-1}}}(N-1)\}.$$

*Theorem 11 (Minimal Spreading Non-Optimality Condition): Under the fixed-size access model and a shifted exponential service time, the minimal spreading allocation does not maximize the DSS service rate $\mu_s(\alpha)$ when $r \geq \min_{2\leq\alpha\leq r}\{\sqrt[\alpha-1]{\frac{\Delta\mu m(m\alpha-\alpha+1)+m\alpha^2}{\alpha(\Delta\mu+1)(m\alpha-\alpha+1)}} \cdot (N-\alpha+1) + \alpha - 1\}$.*

The conditions in Theorems 10 and 11 depend on the number of accessed nodes $r$. For the shifted exponential service, we have Conjecture 1, providing guidelines on finding the optimal $\alpha$ for maximizing $\mu_s(\alpha)$.

*2) Numerical Analysis:* In Fig. 10, we evaluate the expression (21) for $\mu_s(\alpha)$ to see how the DSS service rate changes with $\alpha$. We consider a system with $N = 40$ storage nodes. Using Theorems 10 and 11, we can easily calculate the

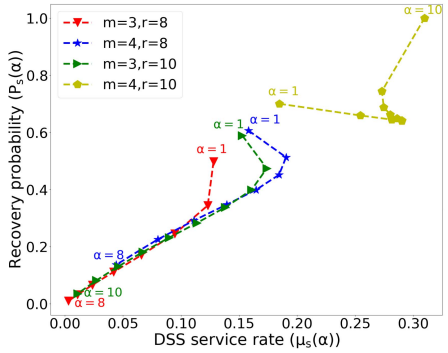



Fig. 11. Successful recovery probability $P_s(\alpha)$ vs. the DSS service rate $\mu_s(\alpha)$ for the fixed-size access model as a function of $\alpha$ for different values of $m$ and $r$ (cf. (11) and (21)). The number of storage nodes is $N = 40$, and the service time follows S-Exp$(3, 1)$. The minimal (or maximal) spreading allocation is optimal when $m$ and $r$ are sufficiently small (or large). Otherwise, the optimal allocation is different for different performance metrics.

optimality and non-optimality conditions. For example, when $m = 2$, the minimal spreading allocation is optimal when $r < 6$ and is non-optimal when $r > 36$. These conditions only provide limited knowledge of the optimal allocation and further insight on the optimal allocation can be found from Fig. 10. The upper graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different redundancy levels $m \in \{1, 2, 3, 4\}$ where the number of accessed nodes is $r = 10$. The lower graph shows $\mu_s(\alpha)$ vs. $\alpha$ for $r \in \{10, 13, 17, 20\}$, and the redundancy level $m = 2$. In the upper subfigure, when $m \leq 2$, the minimal spreading allocation is optimal. When $m = 3$, $\mu_s(\alpha)$ reaches its maximum at $\alpha = 3$. When $m = 4$, the maximal spreading allocation is optimal. In the lower subfigure, when $r \leq 13$, the minimal spreading allocation is optimal, while for $r = 17$, the allocation with $\alpha = 2$ is optimal. When $r = 20$, although the allocation with $\alpha = 4$ is optimal, the maximal spreading allocation provides a local maximum value which is close to the global maximum value. From the observations, we conclude that the minimal spreading allocation is optimal only when $m$ or $r$ is sufficiently small. Otherwise, an allocation with $\alpha \geq 2$ is optimal.

In Fig. 11, we analyze $P_s(\alpha)$ vs. $\mu_s(\alpha)$ as $\alpha$ increases from 1 to $r$. We consider a system with $N = 40$ storage nodes, $m \in \{3, 4\}$, and $r \in \{8, 10\}$. Some observations can be made from the figure: when both $m$ and $r$ are sufficiently small, e.g. $m = 3$ and $r = 8$, the minimal spreading allocation is optimal for both $P_s(\alpha)$ and $\mu_s(\alpha)$. When both $m$ and $r$ are sufficiently large, e.g. $m = 4$ and $r = 10$, the maximal spreading allocation is optimal. Otherwise, e.g. $m = 3$ and $r = 10$, there is no optimal $\alpha$ that maximizes both $P_s(\alpha)$ and $\mu_s(\alpha)$ simultaneously.

### B. Probabilistic Access Model

For the probabilistic access and a shifted exponential service, service rate (9) becomes

$$\mu_s(\alpha) = \sum_{\varphi=\alpha}^{m\alpha} \binom{m\alpha}{\varphi} \frac{\mu\alpha \cdot (1-p)^{\varphi} p^{m\alpha-\varphi}}{\Delta\mu + \alpha(H_{\varphi} - H_{\varphi-\alpha})}. \tag{22}$$

From (22), one can expect that the minimal spreading allocation may not be always optimal. To this end, we present the following result on the optimal $\alpha$ to maximize the service rate when no redundancy is used to store the data.

Assume a constant service time of $\Delta$ for $k$ block. In this case, $\mu_s(\alpha|\varphi(\mathcal{A})) = \alpha/\Delta$ for a set $\mathcal{A}$. Therefore, for no redundancy scenario, i.e. $m = 1$, we have $\mu_s(\alpha) = \frac{\alpha}{\Delta}(1-p)^{\alpha}$.

*Theorem 12 (Optimal $\alpha$): Under the probabilistic access model, a constant service time $\Delta$, and no redundancy case in data storage ($m = 1$), the DSS service rate $\mu_s(\alpha)$ reaches its maximum when $\alpha = \lceil p/(1-p) \rceil$ or $\alpha = \lfloor p/(1-p) \rfloor$.*

*Proof:* Given an integer $\alpha \geq 1$, the DSS service rate $\mu_s(\alpha) = \frac{\alpha}{\Delta}(1-p)^{\alpha}$ is discrete. Then we find the optimal $\alpha$ by comparing the ratio $\mu_s(\alpha)/\mu_s(\alpha+1) = \alpha/(\alpha+1(1-p))$ with 1. Thus, $\frac{\mu_s(\alpha)}{\mu_s(\alpha+1)} \geq 1 \Leftrightarrow \frac{\alpha}{\alpha+1(1-p)} \geq 1 \Leftrightarrow \alpha \leq \frac{p}{1-p}$. Similarly, $\frac{\mu_s(\alpha)}{\mu_s(\alpha+1)} \leq 1 \Leftrightarrow \alpha \geq \frac{p}{1-p}$. Since $\alpha$ is an integer, $\mu_s(\alpha)$ reaches the maximum at $\lceil \frac{p}{1-p} \rceil$ or $\lfloor \frac{p}{1-p} \rfloor$. ■

*Remark 5: Under the probabilistic access model with a constant service time $\Delta$, the minimal spreading allocation ($\alpha = 1$) does not always maximize the DSS service rate.*

Now, we go one step further and find the optimality and non-optimality conditions for the minimal spreading allocation when service time follows a shifted exponential distribution.

*1) Optimality and Non-Optimality Conditions for the Minimal Spreading Allocation:* Considering the complexity of (22), finding the optimal $\alpha$ that maximizes $\mu_s(\alpha)$ in a general scenario is difficult.

*Theorem 13 (Minimal Spreading Optimality Condition): Under the probabilistic access model with a shifted exponential service time, the minimal spreading allocation maximizes the DSS service rate $\mu_s(\alpha)$ when*

$$p \geq \max_{\alpha \geq 2} \left\{ 1 - \sqrt[\alpha-1]{(\Delta\mu + \alpha)/\alpha(\Delta\mu m + 1) \binom{m\alpha - 1}{\alpha - 1}} \right\}.$$

*Proof:* Similar to the proof in Theorem 7. For details, please refer to the supplementary Appendix. ■

*Theorem 14 (Minimal Spreading Non-Optimality Condition): Under the probabilistic access model with a scaled exponential service time, the minimal spreading allocation does not maximize the DSS service rate $\mu_s(\alpha)$ when*

$$p \leq \max_{\alpha \geq 2} \left\{ 1 - \sqrt[\alpha-1]{\frac{m(\Delta\mu(m\alpha - \alpha + 1) + \alpha^2)}{\alpha(\Delta\mu + 1)(m\alpha - \alpha + 1)}} \right\}.$$

*Proof:* Similar to the proof in Theorem 8. For details, please refer to the supplementary Appendix. ■

From Theorems 13 and 14, we see that whether the minimal spreading is definitely optimal or definitely not optimal depends on the probability of failed access $p$. The derived bounds on $p$ that guarantee the optimality or non-optimality of minimal spreading are not tight (cf. Remark 3). However, they help us develop an insight about $\alpha$ that maximizes $\mu_s(\alpha)$, which is stated in Conjecture 2.

*2) Numerical Analysis:* In Fig. 12, we evaluate the expression for $\mu_s(\alpha)$ given in (22) to see how the DSS service rate changes with $\alpha$. We consider a system with $N \geq m\alpha$ storage nodes and the service time follows a shifted exponential distribution with $\Delta = 3$ and $\mu = 1$. Using Theorems 13 and 14, we can easily calculate the optimality and non-optimality conditions. For example, when $m = 2$,

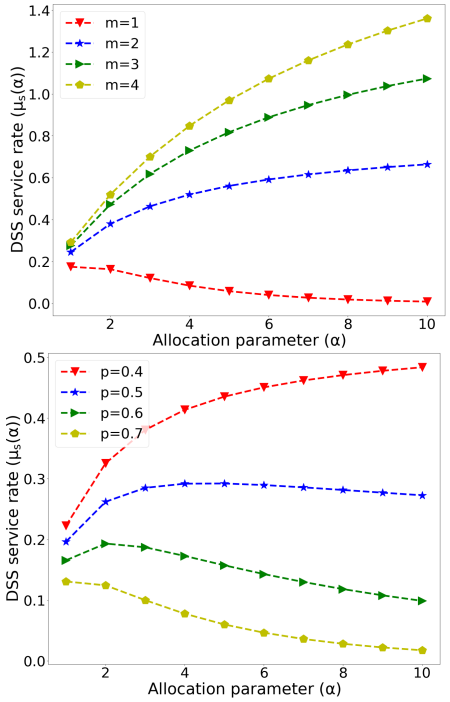



Fig. 12. Service rate $\mu_s(\alpha)$ for the probabilistic access with scaled exponential service as a function of the allocation parameter $\alpha$ (cf. (22)). The number of storage nodes is $N \geq m\alpha$, and the service time follows S-Exp$(3, 1)$. (upper) $\mu_s(\alpha)$ vs. $\alpha$ for the access failure probability $p = 0.3$ and four values of $m$. (lower) $\mu_s(\alpha)$ vs. $\alpha$ for $m = 2$ and four values of $p$. Given $p$ (or $m$), the optimal allocation changes from the minimal spreading allocation to the maximal spreading allocation as $m$ increases or $p$ decreases.

the minimal spreading allocation is optimal when $p > 0.88$ and is non-optimal when $p < 0.08$. These conditions provide only limited knowledge of the optimal allocation and further insight on the optimal allocation can be found from Fig. 12. The upper graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different redundancy levels $m \in \{1, 2, 3, 4\}$, and the failed access probability $p = 0.3$. The lower graph shows $\mu_s(\alpha)$ vs. $\alpha$ for four different failed access probabilities $p \in \{0.4, 0.5, 0.6, 0.7\}$, and the redundancy level $m = 2$. In the upper subfigure, when $m = 1$, the minimal spreading allocation is optimal. When $m \geq 2$, the maximal spreading allocation is optimal. In the lower subfigure, when $p = 0.4$, the maximal spreading allocation is optimal, while for $p = 0.7$, the minimal spreading allocation is optimal. For the other two cases, an allocation with $\alpha \geq 2$ is optimal. As can be seen, the optimal allocation changes with the redundancy level $m$ and the access failure probability $p$.

In Fig. 13, we analyze $P_s(\alpha)$ vs. $\mu_s(\alpha)$ as $\alpha$ increases from 1 to 10. We consider a system with $N \geq m\alpha$ nodes and two values for each parameter: $m \in \{2, 3\}$ and $p \in \{0.45, 0.7\}$. Observe that when $m$ is sufficiently small and $p$ is sufficiently large, the minimal spreading maximizes both $P_s(\alpha)$ and $\mu_s(\alpha)$. On the other hand, for sufficiently large $m$ and sufficiently small $p$, the maximal spreading maximizes both performance metrics. For other cases, there is no optimal $\alpha$ that maximizes both $P_s(\alpha)$ and $\mu_s(\alpha)$ at the same time. For example, when $m = 3$ and $p = 0.7$, $P_s(\alpha)$ reaches its maximum at $\alpha = 1$, and $\mu_s(\alpha)$ reaches its maximum at $\alpha = 10$.

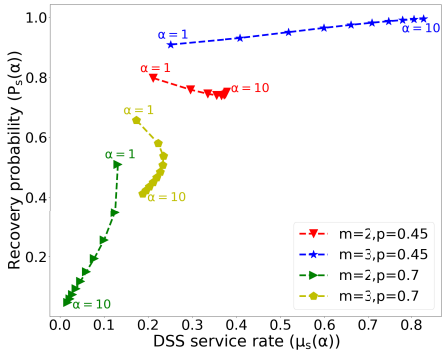



Fig. 13. Successful recovery probability $P_s(\alpha)$ vs. the DSS service rate $\mu_s(\alpha)$ for the probabilistic access model as a function of $\alpha$ for different values of $m$ and $r$ (cf. (13) and (22)). The number of storage nodes is $N \geq m\alpha$, and the service time follows S-Exp$(3, 1)$. The optimal allocation is determined by both $m$ and $r$.

## VII. Conclusion and Future Directions

We considered service rates in distributed storage systems, and focused on two access models (fixed-size and probabilistic access) and two download service models (small and large file). Under the fixed-size access model, a user can access a random fixed-size subset of nodes; under probabilistic access, a user can access each node with a fixed probability. In the small file download model, the randomness associated with the file is negligible; in the large file download model, the randomness is associated with both the file size and inherent system's operations. The primary performance metric of interest is the service rate of the system. Since redundancy for each file is fixed, the allocation of redundancy is essential for improving the system's performance. The general allocation problem is hard to solve. We adopted the common model of quasi-uniform allocation, where coded content is uniformly spread among a subset of storage nodes. Thus the subset size completely specifies the allocation.

Minimal spreading concentrates coded chunks to a minimum-size subset. Maximal spreading allocates coded chunks to each node. For the small file model, the minimal spreading is always optimal. For the large file model, that is not the case. It is not easy to find the optimal allocation. We found the conditions under which the minimal spreading allocation is optimal. We considered scaled exponential and shifted exponential service times. Our numerical results showed that the optimal allocation under these two service models depends on the redundancy level, the number of accessed nodes, and the probability of failed access. As a general rule, one should spread the data blocks to more nodes when the redundancy level is high, the number of accessed nodes is large, or the probability of failed access is small. This work sets the stage for many problems of interest to be studied in the future. We briefly describe three directions of immediate interest.

### A. Optimal Allocations for Non-MDS Codes

Our system model and analysis approach are not limited to MDS codes. If an $[n, k]$ code with minimum distance $d$ is used, then the file can be recovered when any $\ell = n - (d - 1)$ out of $n$ blocks are downloaded. The Singleton bound imposes the constraint $\ell \geq k$, where the equality holds for MDS codes.

When $\ell > k$, the successful recovery condition in (1) becomes $\sum_{i\in\mathcal{A}} s_i \geq \ell$, which means that more storage nodes must be accessed to recover the file. Thus, the exact optimal allocation values we derived may change.

### *B. Optimal Allocations for Other Service Models and Performance Metrics*

We analyzed the most common service time and scaling models. Other distributions, e.g., heavy tail Pareto and Weibull, are also of interest. For large files, it is often appropriate to model download time as the sum of the i.i.d. chunk download times (see [9], [10] for additive service time scaling models). Besides, some other performance metrics, e.g., the expected download time, are also important.

### *C. Minimal Spreading Optimality Conditions*

We found regions of system parameters within which the minimal spreading is optimal. We expect these regions to go beyond the bounds we derived. Finding tighter bounds would help us to better decide on when to use the minimal spreading allocation.

## REFERENCES


[1] M. Noori, E. Soljanin, and M. Ardakani, "On storage allocation for maximum service rate in distributed storage systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 240–244.

[2] P. Peng and E. Soljanin, "On distributed storage allocations of large files for maximum service rate," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 784–791.

[3] G. Joshi, Y. Liu, and E. Soljanin, "Coding for fast content download," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 326–333.

[4] S. Kadhe, E. Soljanin, and A. Sprintson, "Analyzing the download time of availability codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1467–1471.

[5] M. F. Aktas and E. Soljanin, "Straggler mitigation at scale," *IEEE/ACM Trans. Netw.*, vol. 27, no. 6, pp. 2266–2279, Dec. 2019.

[6] M. F. Aktas, S. Kadhe, E. Soljanin, and A. Sprintson, "Download time analysis for distributed storage codes with locality and availability," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3898–3910, Jun. 2021.

[7] G. Yadgar, O. Kolosov, M. F. Aktas, and E. Soljanin, "Modeling the edge: Peer-to-peer reincarnated," in *Proc. 2nd USENIX Workshop Hot Topics Edge Comput., (HotEdge)*, Renton, WA, USA, Jul. 2019, pp. 1–11.

[8] D. Leong, A. G. Dimakis, and T. Ho, "Distributed storage allocations for optimal delay," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 1447–1451.

[9] P. Peng, E. Soljanin, and P. Whiting, "Diversity vs. Parallelism in distributed computing with redundancy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 257–262.

[10] P. Peng, E. Soljanin, and P. Whiting, "Diversity/parallelism trade-off in distributed systems with redundancy," 2020, *arXiv:2010.02147*. [Online]. Available: http://arxiv.org/abs/2010.02147

[11] D. Leong, A. G. Dimakis, and T. Ho, "Distributed storage allocations," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4733–4752, Jul. 2012.

[12] M. Sardari, R. Restrepo, F. Fekri, and E. Soljanin, "Memory allocation in distributed storage networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1958–1962.

[13] B. Hong and W. Choi, "Asymptotic analysis of failed recovery probability in a distributed wireless storage system with limited sum storage capacity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6459–6463.

[14] M. Noori and M. Ardakani, "Allocation for heterogeneous storage nodes," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2102–2105, Dec. 2015.

[15] N. Alon, P. Frankl, H. Huang, V. Rödl, A. Ruciński, and B. Sudakov, "Large matchings in uniform hypergraphs and the conjectures of Erdős and samuels," *J. Combinat. Theory A*, vol. 119, no. 6, pp. 1200–1215, Aug. 2012.

[16] M. Aktas, G. Joshi, S. Kadhe, F. Kazemi, and E. Soljanin, "Service rate region: A new aspect of coded distributed system design," 2020, *arXiv:2009.01598*. [Online]. Available: http://arxiv.org/abs/2009.01598

[17] F. Kazemi, S. Kurz, E. Soljanin, and A. Sprintson, "Efficient storage schemes for desired service rate regions," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.

[18] F. Kazemi, S. Kurz, and E. Soljanin, "A geometric view of the service rates of codes problem and its application to the service rate of the first order Reed-Muller codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 66–71.

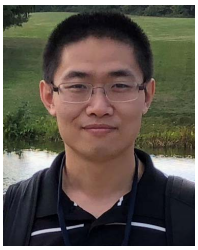
[19] F. Kazemi, E. Karimi, E. Soljanin, and A. Sprintson, "A combinatorial view of the service rates of codes problem, its equivalence to fractional matching and its connection with batch codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 646–651.

[20] S. E. Anderson, A. Johnston, G. Joshi, G. L. Matthews, C. Mayer, and E. Soljanin, "Service rate region of content access from erasure coded storage," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[21] M. Aktas *et al.*, "On the service capacity region of accessing erasure coded content," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 17–24.

[22] Y. Raaijmakers and S. Borst, "Achievable stability in redundancy systems," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 3, pp. 1–21, Nov. 2020.

[23] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. Philadelphia, PA, USA: SIAM, 2008.

**Pei Peng** received the B.S. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2011, and the M.S. degree in electronics and communication engineering from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science, Shanghai, China, in 2014. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department at Rutgers, The State University of New Jersey, USA. During his Ph.D., he has served as a Research and Teaching Assistant. His research interests include coding and allocation in distributed computing systems, covert communications, and machine learning. He has received the 2020 ECE Department Teaching Award.

**Moslem Noori** (Member, IEEE) received the B.Sc. degree in electrical engineering and the B.Sc. degree in applied mathematics from the Amirkabir University of Technology in 2005 and 2006, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Alberta in 2008 and 2012, respectively. From 2013 to 2014, he held a post-doctoral position at The University of British Columbia. He returned to the University of Alberta as an Alberta Innovates Technology Futures Post-Doctoral Fellow from 2014 to 2016. He is currently a Principal Scientist with 1QB Information Technologies. His research interests include distributed storage systems, wireless communications, quantum computation and communication, machine learning for medical applications, and stochastic network analysis. He has received several awards and scholarships, including the NSERC Vanier CGS, the NSERC Post-Doctoral Fellowship, and the AITF Post-Doctoral Fellowship.

**Emina Soljanin** (Fellow, IEEE) is currently a Professor with Rutgers University. Before moving to Rutgers in 2016, she was a (Distinguished) Member of Technical Staff for 21 years in the mathematical sciences research at Bell Labs. Her interests and expertise are wide. Over the past quarter of the century, she has participated in numerous research and business projects, as diverse as power system optimization, magnetic recording, color space quantization, hybrid ARQ, network coding, data and network security, distributed systems performance analysis, and quantum information theory. She served as an Associate Editor for *Coding Techniques* and IEEE TRANSACTIONS ON INFORMATION THEORY, and on the Information Theory Society Board of Governors, and in various roles on other journal editorial boards and conference program committees. She was an Outstanding Alumnus of the Texas A&M School of Engineering, the 2011 Padovani Lecturer, a 2016–2017 Distinguished Lecturer, and the 2019 President of the IEEE Information Theory Society.