Reliability Evaluation of Erasure Coded Systems under Rebuild Bandwidth Constraints

Ilias Iliadis
IBM Research – Zurich
8803 Rüschlikon, Switzerland
Email: ili@zurich.ibm.com

***Abstract*—Modern storage systems employ erasure coding redundancy and recovering schemes to ensure high data reliability at high storage efficiency. The widely used replication scheme belongs to this broad class of erasure coding schemes. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss** (MTTDL) **and the Expected Annual Fraction of Data Loss** (EAFDL) **metrics. To improve the reliability of data storage systems, certain data placement and rebuild schemes reduce the rebuild times by recovering data in parallel from the storage devices. It is often assumed that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. In large-scale data storage systems, however, the network bandwidth is constrained. This article obtains** MTTDL **and** EAFDL **of erasure coded systems analytically for arbitrary rebuild time distributions and for the symmetric, clustered, and declustered data placement schemes under network rebuild bandwidth constraints. The resulting reliability degradation is assessed and the results obtained establish that the declustered placement scheme offers superior reliability in terms of both metrics. Efficient codeword configurations that achieve high reliability in the presence of network rebuild bandwidth constraints are identified.**

Keywords–Storage; Reliability; Data placement; MTTDL; EAFDL; RAID; MDS codes; Information Dispersal Algorithm; Prioritized rebuild; Repair bandwidth; Network bandwidth constraint.

I. INTRODUCTION

In today's large-scale data storage systems, data redundancy is introduced to ensure that data lost due to device and component failures can be recovered. Appropriate redundancy schemes are deployed to prevent permanent loss of data and, consequently, enhance the reliability of storage systems [1]. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss (MTTDL) [2-21] and, more recently, the Fraction of Data Loss Per Year (FDLPY) [22] and the equivalent Expected Annual Fraction of Data Loss (EAFDL) reliability metrics [23-25]. Analytical reliability expressions for MTTDL were obtained predominately using Markovian models, which assume that component failure and rebuild times are independent and exponentially distributed. In practice, however, these distributions are not exponential. To cope with this issue, system reliability was assessed in [17][19][24][25] using an alternative methodology that does not involve Markovian analysis and considers the practical case of non-exponential failure and rebuild time distributions. Moreover, the misconception reported in [26] that MTTDL derivations based on Markovian models provide unrealistic results was dispelled in [27] by invoking improved MTTDL derivations that yield satisfactory results, and also by drawing on prior work that obtains MTTDL analytically without involving Markovian analysis.

Earlier works have predominately considered the MTTDL metric, whereas recent works have also considered the EAFDL metric [23][24][25]. The introduction of the latter metric was motivated by the fact that Amazon S3 considers the durability of data over a given year [28], and, similarly, Facebook [29], LinkedIn [30] and Yahoo! [31] consider the amount of data lost in given periods.

To protect data from being lost and to improve the reliability of data storage systems, replication-based storage systems spread replicas corresponding to data stored on each storage device across several other storage devices. To improve the low storage efficiency associated with the replication schemes, erasure coding schemes that provide a high data reliability as well as a high storage efficiency are deployed. Special cases of such codes are the Redundant Arrays of Inexpensive Disks (RAID) schemes, such as RAID-5 and RAID-6, that have been deployed extensively in the past thirty years [2][3].

State-of-the-art data storage systems [32-35] employ more general erasure codes that affect the reliability, performance, and the storage and reconstruction overhead of the system. In this article, we focus on the reliability assessment of erasure coded systems in terms of the MTTDL and EAFDL metrics. These metrics were analytically derived in [25] for the symmetric, clustered, and declustered data placement schemes under the assumption that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. For instance, in the case of a declustered placement, redundant data associated with the data stored on a given device is placed across all remaining devices in the system. In this way, the rebuild process can be parallelized, which in turn results in short rebuild times. The restoration time can be minimized provided there is sufficient network rebuild bandwidth available. In large-scale erasure coded storage systems, however, the rebuild operations generate a significant amount of network traffic that interferes with user-generated network traffic [36]. A common practice to cope with growing network traffic is to throttle the network bandwidth available for recovery operations, which leads to the network bandwidth being constrained. This in turn results in a reliability degradation, the extent of which is minimized by employing a prioritized rebuild process that first rebuilds the most-exposed to failure data [25][32].

The effect of network rebuild bandwidth constraints on the reliability of replication-based storage systems was studied in [9][16]. It was found that system reliability was significantly

reduced when replicas are spread over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process. The reliability of erasure coded systems in the absence of bandwidth constraints was assessed in [25]. The MTTDL and EAFDL metrics were obtained analytically for the symmetric, clustered, and declustered data placement schemes based on a general framework and methodology. In this article, we recognize that this methodology also holds in the case of network rebuild bandwidth constraints and apply it to derive enhanced closed-form reliability expressions for the MTTDL and EAFDL metrics for these placement schemes in the presence of such rebuild bandwidth constraints. Subsequently, we provide insight into the effect of the placement schemes and the impact of the available network rebuild bandwidth on system reliability. The validity of this methodology for accurately assessing the reliability of storage systems was confirmed by simulations in several contexts [15][16][17][19][23]. It was demonstrated that theoretical predictions for the reliability of systems comprised of highly reliable storage devices are in good agreement with simulation results. Consequently, the emphasis of the present work is on the theoretical assessment of the effect of network rebuild bandwidth constraints on the reliability of erasure coded systems. Moreover, this work extends the reliability results obtained in [16] for the special case of replication-based storage systems to the more general case of erasure coded systems.

The key contributions of this article are the following. We consider the reliability of erasure coded systems under network rebuild bandwidth constraints that was assessed in our earlier work [1] for deterministic rebuild times. In this study, we extend our previous work by also considering arbitrary rebuild times. We show that the codeword lengths that optimize the MTTDL and EAFDL metrics are similar. Furthermore, we derive the asymptotic analytic expressions for the MTTDL and EAFDL reliability metrics when the number of devices becomes large. We then obtain analytically the optimal codeword lengths corresponding to large storage systems. We subsequently establish theoretically that, for large storage systems that use a declustered placement scheme, both metrics are optimized when the codeword length is about 60% of the storage system size, regardless of the rebuild bandwidth constraints.

The remainder of the article is organized as follows. Section II describes the storage system model and the corresponding parameters considered. Section III presents the adaptation of a general framework and methodology for deriving the MTTDL and EAFDL metrics analytically for the case of erasure coded systems under network rebuild bandwidth constraints. Closed-form expressions for the symmetric, clustered, and declustered placement schemes are derived. In Section IV, the data placement schemes that offer the best reliability are identified and the resulting optimal system configurations are specified in Section V. Section VI presents numerical results demonstrating the effectiveness of the erasure coding redundancy schemes for improving system reliability. It also assesses the sensitivity to the network rebuild bandwidth constraints under various codeword configurations. Section VII provides a discussion of the applicability of the results obtained. Finally, we conclude in Section VIII.

II. Storage System Model

Modern data storage systems use erasure coded schemes to protect data from device failures. When devices fail, the redundancy of the affected data is reduced and eventually lost. To avoid irrecoverable data loss, the system performs rebuild operations that use the data stored in the surviving devices to reconstruct the temporarily lost data, thus maintaining the initial data redundancy. We proceed by briefly reviewing the basic concepts of erasure-coding and data-recovery procedures of such storage systems. To assess their reliability, we consider the model used in [25], and adopt and extend the notation. More precisely, the storage system considered here comprises n storage devices (nodes or disks), with each device storing an amount c of data, such that the total storage capacity of the system is $n\,c$.

A. Redundancy

User data is divided into blocks (or symbols) of a fixed size (e.g., sector size of 512 bytes) and complemented with parity symbols to form codewords. We consider (m, l) maximum distance separable (MDS) erasure codes, which are a mapping from l user-data symbols to a set of m ($> l$) symbols, called a codeword, having the property that any subset containing l of the m symbols of the codeword can be used to decode (reconstruct, recover) the codeword. The corresponding storage efficiency s_{eff} is given by

$$s_{\text{eff}} = \frac{l}{m} \,. \tag{1}$$

Consequently, the amount U of user data stored in the system is given by

$$U = s_{\text{eff}}\, n\, c \;=\; \frac{l\, n\, c}{m} \,. \tag{2}$$

Our notation is summarized in Table I. The parameters are divided according to whether they are independent or derived, and are listed in the upper and lower part of the table, respectively.

The m symbols of each codeword are stored on m distinct devices, such that the system can tolerate any $\tilde{r} - 1$ device failures, but $\tilde{r}$ device failures may lead to data loss, with

$$\tilde{r} = m - l + 1 \,. \tag{3}$$

From the above, it follows that

$$1 \le l < m \qquad \text{and} \qquad 2 \le \tilde{r} \le m \,. \tag{4}$$

Examples of MDS erasure codes are the following:

Replication: A replication-based system with a replication factor r can tolerate any loss of up to $r - 1$ copies of some data, such that $l = 1$, $m = r$ and $\tilde{r} = r$. Also, its storage efficiency is equal to $s_{\text{eff}}^{(\text{replication})} = 1/r$.
RAID-5: A RAID-5 array comprised of N devices uses an $(N, N - 1)$ MDS code, such that $l = N - 1$, $m = N$ and $\tilde{r} = 2$. It can therefore tolerate the loss of up to one device, and its storage efficiency is equal to $s_{\text{eff}}^{(\text{RAID-5})} = (N - 1)/N$.
RAID-6: A RAID-6 array comprised of N devices uses an $(N, N - 2)$ MDS code, such that $l = N - 2$, $m = N$ and $\tilde{r} = 3$. It can therefore tolerate a loss of up to two devices, and its storage efficiency is equal to $s_{\text{eff}}^{(\text{RAID-6})} = (N - 2)/N$.
Reed–Solomon: It is based on (m, l) MDS erasure codes.

TABLE I. NOTATION OF SYSTEM PARAMETERS

Parameter	Definition
n	number of storage devices
c	amount of data stored on each device
l	number of user-data symbols per codeword ($l \geq 1$)
m	total number of symbols per codeword ($m > l$)
(m, l)	MDS-code structure
s	symbol size
k	spread factor of the data placement scheme, or group size (number of devices in a group) ($m \leq k \leq n$)
b	average reserved rebuild bandwidth per device
$B_{\max}$	upper limitation of the average network rebuild bandwidth
X	time required to read (or write) an amount c of data at an average rate b from (or to) a device
$F_X(.)$	cumulative distribution function of X
$F_\lambda(.)$	cumulative distribution function of device lifetimes
s_{eff}	storage efficiency of redundancy scheme ($s_{\text{eff}} = l/m$)
U	amount of user data stored in the system ($U = s_{\text{eff}}\, n\, c$)
$\tilde{r}$	minimum number of codeword symbols lost that lead to an irrecoverable data loss ($\tilde{r} = m - l + 1$ and $2 \leq \tilde{r} \leq m$)
N_b	maximum number of devices from which rebuild can occur at full speed in parallel ($N_b = B_{\max}/b$)
ϕ	bandwidth constraint factor $\left(\phi = \min\left(\frac{N_b}{k}, 1\right)\right)$
B_{eff}	effective average network rebuild bandwidth
$f_X(.)$	probability density function of X ($f_X(.) = F'_X(.)$)
μ^{-1}	mean time to read (or write) an amount c of data at an average rate b from (or to) a device ($\mu^{-1} = E(X) = c/b$)
λ^{-1}	mean time to failure of a storage device ($\lambda^{-1} = \int_0^\infty [1 - F_\lambda(t)]dt$)

Note that the RAID-10 and RAID-01 storage systems are non-MDS in that they can sustain a single disk failure and potentially a second one. Similarly, the nested two-dimensional RAID-5 systems, such as RAID 51, use non-MDS erasure codes in that they can sustain any three device failures, but also certain other subsets of more than three device failures [21].

B. Symmetric Codeword Placement

According to a symmetric codeword placement, each codeword is stored on m distinct devices with one symbol per device. In a large storage system, the number of devices n is usually much larger than the codeword length m. Therefore, there are many ways in which a codeword of m symbols can be stored across a subset of the n devices. For each device in the system, the *redundancy spread factor* k denotes the number of devices over which the codewords stored on that device are spread [19]. The system effectively comprises n/k disjoint groups of k devices. Each group contains an amount U/k of user data, with the corresponding codewords placed on the corresponding k devices in a distributed manner. Each codeword is placed entirely in one of the n/k groups. Within each group, all $\binom{k}{m}$ possible ways of placing m symbols across k devices are used equally to store all the codewords in that group.

In such a symmetric placement scheme, within each of the n/k groups, the $m-1$ codeword symbols corresponding to the data on each device are spread *equally* across the remaining $k-1$ devices, the $m-2$ codeword symbols corresponding to the codewords shared by any two devices are spread equally across the remaining $k-2$ devices, and so on. Note also that the n/k groups are logical and therefore need not be physically located in the same node/rack/datacenter.

From the above, it follows that

$$l < m \leq k \leq n\ . \tag{5}$$

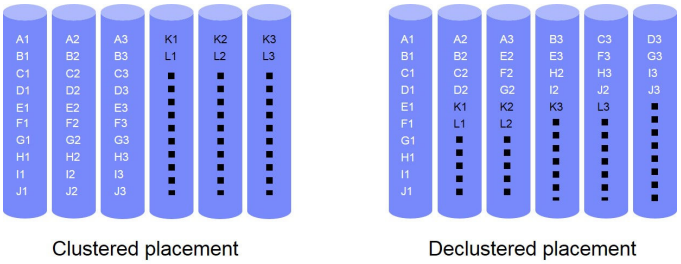


Figure 1. Clustered and declustered placement of codewords of length $m = 3$ on $n = 6$ devices. X1, X2, X3 represent a codeword (X = A, B, C, ... , L).

We proceed by considering the clustered and declustered placement schemes, which are special cases of symmetric placement schemes for which k is equal to m and n, respectively. This results in n/m groups for clustered and one group for declustered placement schemes.

1) Clustered Placement: The n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to the *clustered* placement, each codeword is stored across the devices of a particular cluster, as shown in Figure 1. In such a placement scheme, it can be seen that no cluster stores the redundancies that correspond to the data stored on another cluster. The entire storage system can essentially be modeled as consisting of n/m independent clusters. In each cluster, data loss occurs when $\tilde{r}$ devices fail successively before rebuild operations can be completed successfully.

2) Declustered Placement: In this placement scheme, all $\binom{n}{m}$ possible ways of placing m symbols across n devices are used equally to store all the codewords in the system, as shown in Figure 1.

The clustered and declustered placement schemes represent the two extremes in which the symbols of the codewords associated with the data stored on a failing device are spread across the remaining devices and hence the extremes of the degree of parallelism that can be exploited when rebuilding this data. For declustered placement, the symbols are spread equally across *all* remaining devices, whereas for clustered placement, the symbols are spread across the smallest possible number of devices.

C. Codeword Reconstruction

When storage devices fail, codewords lose some of their symbols, and this leads to a reduction in data redundancy. The system attempts to maintain its redundancy by reconstructing the lost codeword symbols using the surviving symbols of the affected codewords. We assume that failures are detected instantaneously, which immediately triggers the rebuild process.

When a declustered placement scheme is used, as shown in Figure 2, spare space is reserved on each device for temporarily storing the reconstructed codeword symbols before they are transferred to a new replacement device. The rebuild process used to restore the data lost by failed devices is assumed to be both *prioritized* and *distributed*. As discussed in [25], a prioritized (or intelligent) rebuild process always attempts first to rebuild the *most-exposed* codewords, namely, the codewords that have lost the largest number of symbols. The prioritized rebuild process recovers one of the symbols that each of the

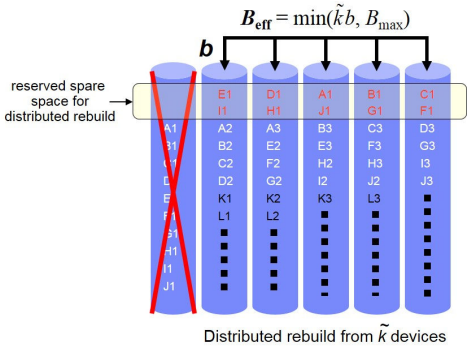


Figure 2. Rebuild under declustered placement.

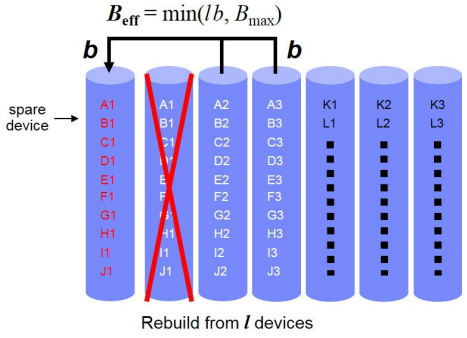


Figure 3. Rebuild under clustered placement.

most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols. In a distributed rebuild process, the codeword symbols lost by failed devices are reconstructed by reading surviving symbols from a number, say $\tilde{k}$, of surviving devices and storing the recovered symbols in the reserved spare space of the $\tilde{k}$ surviving devices, as shown in Figure 2.

A certain proportion of the device bandwidth is reserved for data recovery during the rebuild process, where b denotes the actual average reserved rebuild bandwidth per device. This bandwidth is usually only a fraction of the total bandwidth available at each device, the remaining bandwidth being used to serve user requests. Thus, the lost symbols are rebuilt in parallel using the rebuild bandwidth b available on each surviving device. During this process, it is desirable to reconstruct the lost codeword symbols on devices in which another symbol of the same codeword is not already present. Assuming that the system is at exposure level u (as described in Section II-D below), b_u $(\leq b)$ denotes the average rate at which the amount of data that needs to be rebuilt (repair traffic) is written to selected device(s). Also, denote the cumulative distribution function of the time X required to read (or write) an amount c of data from (or to) a device by $F_X(.)$ and its corresponding probability density function by $f_X(.)$. The kth moment of X, $E(X^k)$, is then given by

$$E(X^k) = \int_0^\infty t^k f_X(t)dt\ , \quad \text{for } k = 1, 2, \dots\ . \tag{6}$$

In particular, $1/\mu$ denotes the average time required to read (or write) an amount c of data from (or to) a device, given by

$$\frac{1}{\mu} \triangleq E(X) = \frac{c}{b}\ . \tag{7}$$

In a distributed rebuild process involving $\tilde{k}$ devices, performing a rebuild at full speed consumes an average network bandwidth of $\tilde{k}\,b$. Let $B_{\max}$ $(\geq b)$ denote the available average network bandwidth for rebuilds. Then, the effective average network rebuild bandwidth used by rebuilds, $B_{\text{eff}}(\tilde{k})$, cannot exceed $B_{\max}$ and is therefore given by

$$B_{\text{eff}}(\tilde{k}) = \min(\tilde{k}\,b, B_{\max}) = \min(\tilde{k}, N_b)\,b\ , \tag{8}$$

where N_b specifies the effective maximum number of devices from which rebuild can occur in parallel at full speed, and is given by

$$N_b \triangleq \frac{B_{\max}}{b}\ . \tag{9}$$

Note that N_b may not be an integer; it only represents the *effective* maximum number of devices from which distributed rebuild can occur at full speed.

A similar reconstruction process is used for other symmetric placement schemes within each group of k devices, except for clustered placement. When clustered placement is used, the codeword symbols are spread across all $k = m$ devices in each group (cluster). Therefore, reconstructing the lost symbols on the surviving devices of a group will result in more than one symbol of the same codeword on the same device. To avoid this, the lost symbols are reconstructed directly in spare devices as shown in Figure 3. In these reconstruction processes, decoding and re-encoding of data are assumed to be done on the fly, so the reconstruction time is equal to the time taken to read and write the required data to the devices. Note also that alternative erasure coding schemes have been proposed to reduce the amount of data transferred over the storage network during reconstruction (see [37][38] and references therein).

D. Exposure Levels and Amount of Data to Rebuild

At time t, $D_j(t)$ denotes the number of codewords that have lost j symbols, where $0 \leq j \leq \tilde{r}$. The system is at exposure level u $(0 \leq u \leq \tilde{r})$, where

$$u = \max_{D_j(t)>0} j. \tag{10}$$

The system is at exposure level u if there are codewords with $m - u$ symbols left, but there are no codewords with fewer than $m - u$ symbols left in the system, that is, $D_u(t) > 0$, and $D_j(t) = 0$, for all $j > u$. These codewords are referred to as the *most-exposed* codewords. At $t = 0$, $D_j(0) = 0$ for all $j > 0$, and $D_0(0)$ is the total number of codewords stored in the system. Device failures and rebuild processes cause the values of $D_1(t), \cdots, D_{\tilde{r}}(t)$ to change over time, and when a data loss occurs, $D_{\tilde{r}}(t) > 0$. Device failures cause transitions to higher exposure levels, whereas rebuilds cause transitions to lower ones. Let t_u denote the time of the first transition from exposure level $u - 1$ to exposure level u, and t_u^+ the instant immediately after t_u. Then, the number C_u of most-exposed codewords when entering exposure level u, $u = 1, \dots, \tilde{r}$, is given by $C_u = D_u(t_u^+)$. For $u = 1$, according to [25, Equation (8)], it holds that $C_1 = c/s$, where s denotes the symbol size. For $u \geq 2$, according to [25, Equations (6) and (27)], the

expected value of C_u is given by

$$E(C_u \,|\, \alpha_1, \ldots, \alpha_{u-1}) = \frac{c}{s} \prod_{j=1}^{u-1} V_j \, \alpha_j \;, \quad \text{for} \;\; u = 2, \ldots, \tilde{r} \;, \tag{11}$$

where V_j represents the fraction of the most-exposed codewords at exposure level j that have symbols stored on a newly failed device that causes the exposure level transition $j \rightarrow j+1$. Note that this fraction is dependent only on the codeword placement scheme. Also, α_j denotes the fraction of rebuild time R_j still left when another device fails causing the exposure level transition $j \rightarrow j+1$. Unconditioning (11) on $\alpha_1, \ldots, \alpha_{u-1}$ yields

$$E(C_u) = \frac{c}{s} \left(\prod_{j=1}^{u-1} V_j \right) \, E \left(\prod_{j=1}^{u-1} \alpha_j \right) \;, \quad \text{for} \;\; u = 2, \ldots, \tilde{r} \;. \tag{12}$$

Analytic expressions for the reliability metrics of interest were derived in [25] using the direct path approximation, which considers only transitions from lower to higher exposure levels [15][17][19]. This implies that each exposure level is entered only once.

E. Failure and Rebuild Time Distributions

We adopt the model and notation considered in [25]. The lifetimes of the n devices are assumed to be independent and identically distributed, with a cumulative distribution function $F_\lambda(.)$ and a mean of $1/\lambda$. We consider real-world distributions, such as Weibull and gamma, as well as exponential distributions that belong to the large class defined in [17]. Note that, although the model considered here does not account for correlated device failures, their effect can be assessed by enhancing the model according to the approach presented in [14]. This issue, however, is beyond the scope of this article. The storage devices are characterized to be *highly reliable* in that the ratio of the mean time $1/\mu$ to read all contents of a device (which typically is on the order of tens of hours), to the mean time to failure of a device $1/\lambda$ (which is typically at least on the order of thousands of hours) is very small, that is,

$$\frac{\lambda}{\mu} = \frac{\lambda \, c}{b} \ll 1 \;. \tag{13}$$

We consider storage devices the cumulative distribution function F_λ satisfies the condition

$$\mu \int_0^\infty F_\lambda(t)[1 - F_X(t)]dt \ll 1, \quad \text{with} \; \frac{\lambda}{\mu} \ll 1 \;, \tag{14}$$

such that the MTTDL and EAFDL reliability metrics of erasure coded storage systems tend to be insensitive to the device failure distribution, that is, they depend only on its mean $1/\lambda$, but not on its density $F_\lambda(.)$. They are, however, sensitive to the distribution $F_X(.)$ of the device rebuild times [25].

III. Derivation of MTTDL and EAFDL

The MTTDL metric assesses the expected amount of time until some data can no longer be recovered and therefore is irrecoverably lost, whereas the EAFDL assesses the fraction of stored data that is expected to be lost by the system annually. The EAFDL is obtained as the ratio of the expected amount of user data lost normalized to the amount of user data to the mean time to data loss [23, Equation (6)]:

$$\text{EAFDL} = \frac{E(H)}{U \cdot \text{MTTDL}} \;, \tag{15}$$

where H denotes the amount of user data lost, given that a data loss has occurred, and with the MTTDL expressed in years.

The MTTDL(B_{max}) and EAFDL(B_{max}) metrics are derived as a function of B_{max} based on the framework and methodology presented in [25]. More specifically, this methodology uses the direct path approximation and does not involve Markovian analysis. It holds for general failure time distributions, which can be exponential or non-exponential, such as the Weibull and gamma distributions that satisfy condition (14). Note that this framework is general because it is also valid in the case where the network rebuild bandwidth is constrained. The only parameters that are affected by the network rebuild bandwidth constraint are the rebuild rates and, accordingly, those parameters that are dependent on them, such as the rebuild times. Analytic expressions for the two metrics of interest were derived in [25, Equations (49) and (50)] as follows:

$$\text{MTTDL}(B_{\text{max}}) \approx \frac{1}{n\,\lambda} \, \frac{(\tilde{r}-1)!}{(\lambda\, c)^{\tilde{r}-1}} \, \frac{[E(X)]^{\tilde{r}-1}}{E(X^{\tilde{r}-1})} \, \prod_{u=1}^{\tilde{r}-1} \frac{b_u(B_{\text{max}})}{\tilde{n}_u} \, \frac{1}{V_u^{\tilde{r}-1-u}} \;, \tag{16}$$

and

$$\text{EAFDL}(B_{\text{max}}) \approx m\,\lambda\,(\lambda\, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \, \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}} \, \prod_{u=1}^{\tilde{r}-1} \frac{\tilde{n}_u}{b_u(B_{\text{max}})} \, V_u^{\tilde{r}-u} \;, \tag{17}$$

where $\tilde{n}_u$ represents the number of devices at exposure level u whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $u+1$. As mentioned above, b_u, the average rate at which the amount of data that needs to be rebuilt at exposure level u is written to selected device(s), is dependent on B_{max}, the upper limitation of the average network rebuild bandwidth.

The expected amount $E(Q)$ of data lost upon a first-device failure is given by [25, Equation (47)]

$$E(Q) \approx \, l\, c\, (\lambda\, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \, \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}} \, \prod_{u=1}^{\tilde{r}-1} \frac{\tilde{n}_u}{b_u} \, V_u^{\tilde{r}-u} \;, \tag{18}$$

where $E(X^{\tilde{r}-1})$ is obtained by (6). Subsequently, the expected amount $E(H)$ of data lost, given that a data loss has occurred, is given by [25, Equation (48)]

$$E(H) \approx \left(\frac{l}{\tilde{r}} \, \prod_{u=1}^{\tilde{r}-1} V_u \right) c \;. \tag{19}$$

Central to the derivation of $E(Q)$ and $E(H)$ is the assessment of the amount H of user data lost, that is, the amount of user data stored in the most-exposed codewords when entering exposure level $\tilde{r}$ that can no longer be recovered and therefore is irrecoverably lost. In [25, Equation (22)] it was assumed that

each of the most-exposed codewords loses all its l symbols of user data, that is,

$$H = C_{\tilde{r}}\, l\, s\ , \tag{20}$$

where $C_{\tilde{r}}$ is the number of the most-exposed codewords when entering exposure level $\tilde{r}$, and s is the symbol size. This clearly overestimates the amount of data lost, especially when the number of user-data symbols l in each of the most-exposed codewords is greater than the number of lost symbols $\tilde{r}$. We proceed to revise the derivation of H. Let n_l denote the number of user-data symbols irrecoverably lost in a typical most-exposed codeword. As devices are equally likely to fail, and given that the fraction of user-data symbols in a codeword is equal to l/m, a moment's reflection reveals that the expected number of user-data symbols irrecoverably lost is given by

$$E(n_l) = \frac{l}{m}\,\tilde{r}\ =\ \frac{\tilde{r}}{m}\,l\ . \tag{21}$$

Consequently, for a number $C_{\tilde{r}}$ of most-exposed codewords, the expected amount $E(H\,|\,C_{\tilde{r}})$ of user data lost is given by

$$E(H\,|\,C_{\tilde{r}}) = C_{\tilde{r}}\,E(n_l)\,s\ \overset{(21)}{=}\ C_{\tilde{r}}\,\frac{\tilde{r}}{m}\,l\,s\ , \tag{22}$$

which in turn, by unconditioning on $C_{\tilde{r}}$, yields

$$E(H) = E(C_{\tilde{r}})\,\frac{\tilde{r}}{m}\,l\,s\ . \tag{23}$$

From the above, it follows that in each of the most-exposed codewords, the expected fraction f_l of the user-data symbols that are lost is given by

$$f_l\ \triangleq\ E\left(\frac{n_l}{l}\right) = \frac{E(n_l)}{l} = \frac{\tilde{r}}{m}\ \overset{(3)}{=}\ \frac{m-l+1}{m}\ . \tag{24}$$

The resulting expressions for $E(Q)$, EAFDL, and $E(H)$ can now be obtained by multiplying (18), (17), and (19) with f_l, which yields

$$E(Q) \approx\ \frac{l}{m}\,c\,(\lambda\,c)^{\tilde{r}-1}\ \frac{1}{(\tilde{r}-1)\,!}\ \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\ \prod_{u=1}^{\tilde{r}-1}\frac{\tilde{n}_u}{b_u}\ V_u^{\tilde{r}-u}\ , \tag{25}$$

where $E(X^{\tilde{r}-1})$ is obtained by (6),

$$\text{EAFDL}(B_{\max}) \approx$$
$$\lambda\,(\lambda\,c)^{\tilde{r}-1}\ \frac{1}{(\tilde{r}-1)\,!}\ \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\ \prod_{u=1}^{\tilde{r}-1}\frac{\tilde{n}_u}{b_u(B_{\max})}\ V_u^{\tilde{r}-u}\ , \tag{26}$$

and

$$E(H) \approx \left(\frac{l}{m}\ \prod_{u=1}^{\tilde{r}-1} V_u\right) c\ . \tag{27}$$

Remark 1: From (16), (26), and (27), and given that $E(X) = c/b$, it follows that MTTDL and EAFDL are dependent on the $(m-l)$th moment of the rebuild time distribution. Furthermore, given that $E(X^{m-l}) \geq [E(X)]^{m-l}$, random rebuild times result in lower MTTDL and higher EAFDL values than deterministic rebuild times do. In contrast, the expected amount $E(H)$ of user data lost, given that a data loss has occurred, is not dependent on λ, b and c nor on the rebuild time distribution. Moreover, $E(H)$ is not dependent on b_u and therefore is not affected by the limitation on the network rebuild bandwidth; it is only dependent on the storage efficiency and data placement scheme. Moreover, MTTDL is dependent on n, but EAFDL and $E(H)$ are not.

Remark 2: The analytic expressions for the MTTDL and EAFDL reliability metrics were derived in [25] in the absence of network rebuild bandwidth constraints. Consequently, they correspond to the case of $B_{\max} = \infty$, where the two metrics are denoted by MTTDL(∞) and EAFDL(∞), respectively.

From (16) and (17), or the enhanced expression (26), it follows that

$$\frac{\text{MTTDL}(B_{\max})}{\text{MTTDL}(\infty)} = \frac{\text{EAFDL}(\infty)}{\text{EAFDL}(B_{\max})} = \theta\ , \tag{28}$$

where θ represents the *reliability reduction factor* that assesses the reliability degradation due to a network rebuild bandwidth constraint, and is given by

$$\theta\ \triangleq\ \prod_{u=1}^{\tilde{r}-1}\frac{b_u(B_{\max})}{b_u(\infty)}\ . \tag{29}$$

Equation (28) suggests that the reliability reduction factor for EAFDL is the same as the one for MTTDL. At first glance, and given the different nature of the MTTDL and EAFDL metrics, this seems to be counterintuitive. The reason for this result is that network rebuild bandwidth constraints effectively prolong the duration of the rebuild times, which equally affects the MTTDL and EAFDL metrics. More specifically, at exposure level u, and according to [25, Equations (43) and (44)], the transition probability $P_{u\rightarrow u+1}$ from exposure level u to $u+1$ is proportional to the rebuild time R_u, which in turn is inversely proportional to the average rebuild rate b_u. Thus, constraining the average rebuild rates b_u increases the probability of data loss P_{DL}. Consequently, the reliability metrics are equally affected given that, according to [25, Equations (14) and (15)], MTTDL and EAFDL are inversely proportional and proportional to P_{DL}, respectively. Note also that the corresponding amount of data lost H is dependent only on the data placement scheme and is therefore not affected by the prolongation of the rebuild times.

Remark 3: From (29), and given that $b_u(B_{\max})$ decreases with decreasing $B_{\max}$, it follows that θ decreases with increasing $\tilde{r}$ or decreasing $B_{\max}$.

Remark 4: From (12), (23), and (27), it follows that

$$E\left(\prod_{j=1}^{\tilde{r}-1}\alpha_j\right) = \frac{1}{\tilde{r}}\ . \tag{30}$$

Note that the variables $\alpha_1,\ldots,\alpha_{\tilde{r}-1}$ are generally independent and approximately uniformly distributed between 0 and 1 such that $E(\alpha_u) \approx 1/2$, $u = 1,\ldots,\tilde{r}-1$ [23, 25]. In this context, however, this assumption would lead to the erroneous result $E(\prod_{j=1}^{\tilde{r}-1}\alpha_j) = 1/2^{\tilde{r}-1}$, which is smaller than the correct one by a factor of $2^{\tilde{r}-1}/\tilde{r}$. This is analogous to the factor of $2^{r-1}/r$ that was derived for replication-based storage systems in Section V.E of [23]. It turns out that when a data loss has occurred, the variables $\alpha_1,\ldots,\alpha_{\tilde{r}-1}$ are not distributed identically. Further insight regarding this subtle issue is provided in the relevant discussion of that section.

Assuming that the system has reached exposure level u, we deduce from (30) that

$$E\left(\prod_{j=1}^{u-1}\alpha_j\right)=\frac{1}{u}\,,\quad \text{for}\ \ u=2,\ldots,\tilde{r}\,. \tag{31}$$

A. Symmetric Placement

We consider the case where the redundancy spread factor k is in the interval $m<k\leq n$. The special case $k=m$, which corresponds to the clustered placement scheme, has to be considered separately for the reasons discussed in Section II-B1. As discussed in [25, Section IV-B], the *prioritized* rebuild process at each exposure level u recovers one of the u symbols that each of the most-exposed codewords has lost by reading $m-\tilde{r}+1$ of the remaining symbols from the $\tilde{n}_u$ surviving devices in the affected group. According to [25, Equation (51)], it holds that

$$\tilde{n}_u^{\text{sym}}=k-u\,. \tag{32}$$

Furthermore, in the absence of a network rebuild bandwidth constraint, the total write bandwidth, which is also the average rebuild rate b_u, is given by [25, Equation (52)]

$$b_u^{\text{sym}}(\infty)=\frac{\tilde{n}_u^{\text{sym}}}{m-\tilde{r}+2}\,b\overset{(3)}{=}\frac{\tilde{n}_u^{\text{sym}}\,b}{l+1}\,,\quad u=1,\ldots,\tilde{r}-1\,. \tag{33}$$

However, in the presence of a network rebuild bandwidth constraint B_{max} and according to (8) with $\tilde{k}=\tilde{n}_u=\tilde{n}_u^{\text{sym}}$, the average rebuild rate b_u is given as a function of B_{max} by

$$b_u^{\text{sym}}(B_{\text{max}})=\frac{B_{\text{eff}}(\tilde{n}_u)}{l+1}=\frac{\min(\tilde{n}_u\,b,B_{\text{max}})}{l+1}=\frac{\min(\tilde{n}_u,N_b)\,b}{l+1}\overset{(32)}{=}\frac{\min(k-u,N_b)\,b}{l+1}\,,\ \text{for}\ \ u=1,\ldots,\tilde{r}-1\,. \tag{34}$$

Substituting (33) and (34) into (29) yields

$$\theta^{\,\text{sym}}=\prod_{u=1}^{\tilde{r}-1}\frac{\min(k-u,N_b)}{k-u}\,. \tag{35}$$

Note that when $N_b\geq k-1$, the system reliability is not affected because all rebuilds are performed at full speed, and therefore the factor θ is equal to 1. However, when $N_b<k-1$, it may not be possible for some of the rebuilds to be performed at full speed, and therefore the factor θ will be less than 1, which affects the system reliability. Consequently, the reliability reduction factor θ depends on the *bandwidth constraint factor* ϕ which is given by

$$\phi\triangleq\min\left(\frac{N_b}{k},1\right)\overset{(9)}{=}\min\left(\frac{B_{\text{max}}}{k\,b},1\right)\,,\ \text{with}\ 0\leq\phi\leq 1\,. \tag{36}$$

From (34), (35), and (36), and recognizing that $\min(k-u,N_b)=\min(\min(k-u,k),N_b)=\min(k-u,\min(k,N_b))=\min(k\min(1,N_b/k),k-u)=\min(k\,\phi,k-u)$, it follows that

$$b_u^{\text{sym}}(B_{\text{max}})=\frac{\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)(k-u)\,b}{l+1}\,,\ \text{for}\ \ u=1,\ldots,\tilde{r}-1\,, \tag{37}$$

and

$$\theta^{\,\text{sym}}=\prod_{u=1}^{\tilde{r}-1}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)\,. \tag{38}$$

Using (38) and the fact that MTTDL(∞) is given by [25, Equation (54)], (28) yields

$$\text{MTTDL}_k^{\text{sym}}(B_{\text{max}})\approx\frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)!\,\frac{[E(X)]^{m-l}}{E(X^{m-l})}\prod_{u=1}^{m-l}\left(\frac{k-u}{m-u}\right)^{m-l-u}\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)\,, \tag{39}$$

where B_{max} is expressed via ϕ given by (36).

Note that, for a deterministic rebuild time distribution, for which it holds that $E(X^{m-l})=[E(X)]^{m-l}$, and for a replication-based system, for which $m=r$ and $l=1$, and by virtue of (35) and (38), Equation (39) is in agreement with Equation (24) of [16], where $c/b=1/\mu$.

Using (38) and the fact that EAFDL(∞) is obtained by multiplying [25, Equation (55)] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost, by virtue of (24), (28) yields

$$\text{EAFDL}_k^{\text{sym}}(B_{\text{max}})\approx\lambda\left[\frac{(l+1)\,\lambda\,c}{b}\right]^{m-l}\frac{1}{(m-l)!}\,\frac{E(X^{m-l})}{[E(X)]^{m-l}}\prod_{u=1}^{m-l}\left(\frac{m-u}{k-u}\right)^{m-l+1-u}\Bigg/\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right), \tag{40}$$

where B_{max} is expressed via ϕ given by (36).

Moreover, $E(H)$ can be obtained by multiplying [25, Equation (56)] with f_l, which, according to (24), yields

$$E(H)_k^{\text{sym}}\approx\left(\frac{l}{m}\prod_{u=1}^{m-l}\frac{m-u}{k-u}\right)c \tag{41}$$

$$=\frac{l\,(m-1)!\,(k-m+l-1)!}{m\,(k-1)!\,(l-1)!}\,c\,. \tag{42}$$

For given l, m, n, and ϕ, the redundancy spread factors or, equivalently, the optimal group sizes that maximize MTTDL$^{\text{sym}}$, EAFDL$^{\text{sym}}$, and $E(H)^{\text{sym}}$ are given by the following propositions.

Proposition 1: For given l, m, n, and ϕ, and for any λ, c, b, and rebuild time distribution of X, the value of k $(m+1\leq k\leq n)$, denoted by $\hat{k_s}$, that maximizes MTTDL$_k^{\text{sym}}$ is given by

$$\hat{k_s}=\begin{cases}\text{any } j\in[m+1,n]\cap D_n\,,\ \text{which includes } j=k_m\,,\\ \qquad\text{for } m-l=1 \text{ and } \phi\geq 1-\frac{1}{n}\\ \text{any } j\in[m+1,\frac{1}{1-\phi}]\cap D_n\,,\ \text{which includes } j=k_m\,,\\ \qquad\text{for } m-l=1 \text{ and } 1-\frac{1}{k_m}\leq\phi<1-\frac{1}{n}\\ k_m\,,\qquad\text{for } m-l=1 \text{ and } \phi<1-\frac{1}{k_m}\\ n\,,\qquad\text{for } m-l\geq 2\,,\end{cases} \tag{43}$$

where D_n is the set of the divisors (factors) of n, that is,

$$D_n \triangleq \{\, j \,:\, j \,|\, n \} \;\equiv\; \left\{ j \,:\, j \in \mathbb{N} \quad \text{and} \quad \frac{n}{j} \in \mathbb{N} \right\} , \quad (44)$$

and k_m is the smallest integer in the interval $I_k = [m+1, n]$ that divides n, that is,

$$k_m \triangleq \min_j \{\, j \in I_k \cap D_n \} . \quad (45)$$

Proof: See Appendix A. ■

Proposition 2: For given l, m, n, and ϕ, and for any c, λ, and rebuild time distribution of X, $\text{EAFDL}_k^{\text{sym}}$ and $E(H)_k^{\text{sym}}$ are decreasing in k and are therefore minimized when $k = n$.

Proof: Considering l, m, and n to be fixed, it follows from (40) that $\text{EAFDL}_k^{\text{sym}}$ is inversely proportional to the function B_k given by

$$B_k \triangleq \prod_{u=1}^{m-l} (k-u)^{m-l-u} \;\; \min(k\,\phi,\, k-u) . \quad (46)$$

Note that each of the terms in the product is increasing in k, which implies that B_k is also increasing in k and, consequently, $\text{EAFDL}_k^{\text{sym}}$ is decreasing in k. Furthermore, it follows from (41) that $E(H)_k^{\text{sym}}$ is also decreasing in k. ■

Remark 5: From the preceding two propositions it follows that, for $l+1 < m < k \leq n$, $\text{MTTDL}_k^{\text{sym}}$ is maximized and $\text{EAFDL}_k^{\text{sym}}$ and $E(H)_k^{\text{sym}}$ are minimized by the declustered placement scheme, that is, when $k = n$.

An approximate expression for the reliability reduction function is given by the following lemma.

LEMMA 1: For large values of k, m, l, and $m-l$, $\theta^{\text{sym}}_{\text{approx}}$ can be approximated as follows:

$$\log\left(\theta^{\text{sym}}_{\text{approx}}\right) \approx \left[\log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k - \frac{1}{2}\log(1-\widehat{\phi}) , \quad (47)$$

where $\widehat{\phi}$ is given by

$$\widehat{\phi} \;\triangleq\; \min(1-\phi\,, hx) , \quad (48)$$

h is given by

$$h \;\triangleq\; 1 - s_{\text{eff}} = 1 - \frac{l}{m} , \quad (49)$$

and x by

$$x \triangleq \frac{m}{k} . \quad (50)$$

Proof: See Appendix B. ■

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 3: For large values of n, k, m, l, and $m-l$, $\text{MTTDL}^{\text{sym}}$ normalized to $1/\lambda$ can be approximated as follows:

$$\begin{aligned}\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max})\right) \approx\; & \log\left(\frac{k}{n}\right) \\ & + k^2\,\frac{W(h,x)}{2} + k\left\{hx\,\log\left(\frac{hx\sqrt{x}\,k\,b}{e\,[(1-h)x\,k+1]\,\lambda\,c}\right)\right. \\ & \left. + \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right\} \\ & - \frac{1}{8}\left[h(1-x) - \log\left(\frac{1-h}{1-hx}\right)\right] + \log\left(\sqrt{\frac{2\pi hx}{k}}\right) \\ & - \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{[E(X)]^{hxk}}{E(X^{hxk})}\right) , \end{aligned} \quad (51)$$

where $W(h,x)$ is given by

$$W(h,x) \triangleq hx(1-x) - \log\left(\frac{\left[(1-h)^{(1-h)^2}\,x^{h^2}\right]^{x^2}}{(1-hx)^{(1-hx)^2}}\right) , \quad (52)$$

and h, x, and $\widehat{\phi}$ are given by (49), (50), and (48), respectively.

Proof: It follows from (28) that

$$\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max}) \;=\; \text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\;\theta^{\,\text{sym}}_{\text{approx}} , \quad (53)$$

or, equivalently,

$$\begin{aligned}&\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max})\right) = \\ &\qquad \log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\right) + \log\left(\theta^{\,\text{sym}}_{\text{approx}}\right) . \end{aligned} \quad (54)$$

Substituting the analytic expression obtained in [25, Equation (62)] for the term $\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\right)$, and (47) into (54) yields (51). ■

Proposition 4: For large values of k, m, l, and $m-l$, $\text{EAFDL}^{\text{sym}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max})/\lambda\right) \approx\; & -k^2\,\frac{W(h,x)}{2} \\ & + k\left\{hx\,\log\left(\frac{e\,[(1-h)x\,k+1]\,\lambda\,c}{h\,\sqrt{x}\,k\,b}\right) + \log\left(\frac{(1-hx)^{1-hx}}{(1-h)^{(1-h)x}}\right)\right. \\ & \left. - \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) - \widehat{\phi}\right\} \\ & + \frac{1}{8}\,h(1-x) + \log\left(\sqrt{\frac{1}{2\pi hxk}}\left(\frac{1-h}{1-hx}\right)^{\frac{3}{8}}\right) \\ & + \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{E(X^{hxk})}{[E(X)]^{hxk}}\right) , \end{aligned} \quad (55)$$

where $B_{\max}$ is expressed via ϕ given by (36), and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (50), (52), and (48), respectively.

Proof: It follows from (28) that

$$\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max}) \;=\; \frac{\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)}{\theta^{\,\text{sym}}_{\text{approx}}} . \quad (56)$$

or, equivalently,

$$\begin{aligned}&\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max})/\lambda\right) = \\ &\qquad \log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right) - \log\left(\theta^{\,\text{sym}}_{\text{approx}}\right) . \end{aligned} \quad (57)$$

The term $\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)$ is obtained by multiplying the corresponding term obtained in [25] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost. Consequently, the term $\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right)$ is obtained by adding to the analytic expression obtained in [25, Equation (64)] the term $\log(f_l)$, which, according to (24), and using (49) and (50), is given by

$$f_l = \frac{hxk+1}{xk} \;, \quad \text{or} \quad \log(f_l) = \log\left(\frac{hxk+1}{xk}\right) \;. \tag{58}$$

Substituting the outcome for the resulting enhanced term $\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right)$ and (47) into (57) yields (55). ■

Proposition 5: For large values of k, m, l, and $m-l$, $E(H)^{\text{sym}}$ normalized to c can be approximated as follows:

$$\log\left(E(H)^{\text{sym}}_{\text{approx}}/c\right) \approx \log\left((1-h)\sqrt{\frac{1-h}{1-hx}}\right) + k\,V(h,x) \;, \tag{59}$$

where $V(h,x)$ is given by

$$V(h,x) \triangleq \log\left(\frac{x^x\,(1-hx)^{1-hx}}{[(1-h)x]^{(1-h)x}}\right) \;, \tag{60}$$

and h and x are given by (49) and (50), respectively.

Proof: Immediate by multiplying the term $\text{EAFDL}^{\text{sym}}_{\text{approx}}$ with f_l or, equivalently, by adding to [25, Equation (58)] the term $\log(f_l)$ given by (58). ■

B. Clustered Placement

In the clustered placement scheme, the n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to *clustered* placement, each codeword is stored across the devices of a particular cluster. At each exposure level u, the rebuild process recovers one of the u symbols that each of the C_u most-exposed codewords has lost by reading $m-\tilde{r}+1$ of the remaining symbols. Note that the remaining symbols are stored on the $m-u$ surviving devices in the affected group. According to [25, Equation (65)], it holds that

$$\tilde{n}^{\text{clus}}_u = m-u \;. \tag{61}$$

In the case of clustered placement, the rebuild process recovers the lost symbols by reading l symbols from l of the $\tilde{n}_u$ surviving devices of the affected cluster. In the absence of a network rebuild bandwidth constraint, the symbols are read from each of the l devices at an average rate of b such that the average effective network rebuild bandwidth is equal to $B_{\text{eff}} = l\,b$. Subsequently, the lost symbols are computed on-the-fly and written to a spare device at an average rate of $B_{\text{eff}}/l = b$. Consequently, it holds that

$$b^{\text{clus}}_u(\infty) = b \,, \quad u = 1,\ldots,\tilde{r}-1 \;. \tag{62}$$

However, in the presence though of a network rebuild bandwidth constraint $B_{\max}$ the effective average network rebuild bandwidth is equal to $B_{\text{eff}} = \min(l\,b, B_{\max})$, which implies that the lost symbols are written to a spare device at an average rate of B_{eff}/l. Thus, the average rebuild rate b_u is given as a function of $B_{\max}$ by

$$b^{\text{clus}}_u(B_{\max}) = \frac{B_{\text{eff}}(B_{\max})}{l} = \frac{\min(l\,b, B_{\max})}{l} = \frac{\min(l, N_b)\,b}{l} \;, \quad \text{for } u = 1,\ldots,\tilde{r}-1 \;. \tag{63}$$

Remark 6: Note that, as far as the data placement is concerned, the clustered placement scheme is a special case of a symmetric placement scheme for which k is equal to m. However, its reliability assessment cannot be directly obtained from the reliability results derived in Section III-A for the symmetric placement scheme by simply setting $k=m$. The reason for that is the difference in the rebuild processes. In the case of a symmetric placement scheme, recovered symbols are written to the spare space of existing devices, whereas in the case of a clustered placement scheme, recovered symbols are written to a spare device. This results in different rebuild bandwidths, which are given by (34) and (63), respectively.

Substituting (62) and (63) into (29) yields

$$\theta^{\,\text{clus}} = \left(\frac{\min(l, N_b)}{l}\right)^{\tilde{r}-1} \;. \tag{64}$$

As $l < m$, it holds that $\min(l, N_b) = \min(\min(l,m), N_b) = \min(\min(N_b, m), l) = \min(m\min(N_b/m, 1), l) = \min(m\phi, l)$, where, analogously to (36), and with $k=m$,

$$\phi \triangleq \min\left(\frac{N_b}{m}, 1\right) \overset{(9)}{=} \min\left(\frac{B_{\max}}{m\,b}, 1\right) \;, \quad \text{where } 0 \le \phi \le 1 \;. \tag{65}$$

Consequently, (63) and (64) yield

$$b^{\text{clus}}_u(B_{\max}) = \min\left(\frac{m}{l}\,\phi\,, 1\right) b \;, \quad \text{for } u = 1,\ldots,\tilde{r}-1 \;, \tag{66}$$

and

$$\theta^{\,\text{clus}} = \min\left(\frac{m}{l}\,\phi\,, 1\right)^{\tilde{r}-1} \;, \tag{67}$$

respectively.

Remark 7: It follows from (67) that for $m\,\phi/l \ge 1$ or, equivalently, for $\phi \ge s_{\text{eff}} = l/m$, $\theta^{\,\text{clus}}$ is equal to 1, which implies that the bandwidth constraint does not affect the system reliability.

Using (3) and (67), and the fact that $\text{MTTDL}(\infty)$ is given by [25, Equation (68)], (28) yields

$$\text{MTTDL}^{\text{clus}}(B_{\max}) \approx \frac{1}{n\,\lambda}\left(\frac{\min(\frac{m\,\phi}{l}, 1)\,b}{\lambda\,c}\right)^{m-l} \frac{1}{\binom{m-1}{l-1}}\,\frac{[E(X)]^{m-l}}{E(X^{m-l})} \;, \tag{68}$$

where $B_{\max}$ is expressed via ϕ given by (65).

Note that for a RAID-5 array comprised of N devices, such that $n=k=m=N$ and $l=N-1$, for $\phi \ge s_{\text{eff}} = l/m$ and by virtue of (7), (68) yields

$$\text{MTTDL}^{\text{clus}}_{\text{RAID-5}} \approx \frac{\mu}{N(N-1)\lambda^2} \;, \tag{69}$$

which is the same result as that reported in [2]. Note that when $m-l=1$, MTTDL is insensitive to the rebuild time distribution.

For an exponential rebuild time distribution, for which it holds that $E(X^{m-l}) = (m-l)!\,[E(X)]^{m-l}$, and for a RAID-6 array comprised of N devices, such that $n = k = m = N$ and $l = N-2$, for $\phi \geq s_{\text{eff}} = l/m$ and by virtue of (7), (68) yields

$$\text{MTTDL}^{\text{clus, exp}}_{\text{RAID-6}} \;\approx\; \frac{\mu^2}{N(N-1)(N-2)\lambda^3}\;, \tag{70}$$

which is the same result as that reported in [3]. That result was derived using a continuous-time Markov chain (CTMC) model with the repair rate equal to μ, which is not dependent on the number of failed devices. This is analogous to our model where lost symbols are written to a spare device at an average rate of b, which is fixed and is not dependent on the number of failed devices, and the rebuild time distribution is exponential.

In contrast, in [39], the Mean Time Between Failures (MTBF) was derived using a CTMC model and assuming that the repair rate of each failed device is fixed, which implies that the total repair rate is proportional to the number of failed devices. In the case where $\lambda \ll \mu$, [39, Equation (1.1)] with k replaced by l and n by N yields

$$\text{MTBF} \;\approx\; \frac{1}{l\,\lambda\,\binom{N}{l}\,(\lambda/\mu)^{N-l}}\;. \tag{71}$$

For a deterministic rebuild time distribution, for which it holds that $E(X^{m-l}) = [E(X)]^{m-l}$, for $\phi \geq s_{\text{eff}} = l/m$ and for a RAID-6 array, (68) yields

$$\text{MTTDL}^{\text{clus, det}}_{\text{RAID-6}} \;\approx\; \frac{2\,\mu^2}{N(N-1)\lambda^3}\;, \tag{72}$$

and for arbitrary l values ($l < m = k = n = N$), (68) yields

$$\begin{aligned}\text{MTTDL}^{\text{clus, det}} \;&\approx \frac{1}{N\,\lambda}\,\frac{1}{(\lambda/\mu)^{N-l}}\,\frac{1}{\binom{N-1}{l-1}}\\ &= \frac{1}{l\,\lambda\,\binom{N}{l}\,(\lambda/\mu)^{N-l}}\;,\end{aligned} \tag{73}$$

which is the same result as in (71), despite some strikingly different characteristics in the operation of the underlying systems. MTBF is obtained assuming exponential rebuild times (Markovian behavior) and repair rates proportional to the number of failed devices, whereas our model yields the same result assuming deterministic rebuild times (non-Markovian behavior) and a fixed repair rate independent of the number of failed devices.

We now proceed to derive EAFDL and $E(H)$. Using (3) and (38), and the fact that EAFDL(∞) is obtained by multiplying [25, Equation (69)] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost, by virtue of (24), (28) yields

$$\begin{aligned}&\text{EAFDL}^{\text{clus}}(B_{\max}) \approx\\ &\qquad \lambda\left(\frac{\lambda\,c}{\min(\frac{m\,\phi}{l},1)\,b}\right)^{m-l}\binom{m-1}{l-1}\,\frac{E(X^{m-l})}{[E(X)]^{m-l}}\;,\end{aligned} \tag{74}$$

where $B_{\max}$ is expressed via ϕ given by (65).

Moreover, $E(H)$ can be obtained by multiplying [25, Equation (70)] with f_l, which, according to (24), yields

$$E(H)^{\text{clus}} \approx \frac{l}{m}\,c\,. \tag{75}$$

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 6: For large values of n, m, l, and $m-l$, $\text{MTTDL}^{\text{clus}}$ normalized to $1/\lambda$ and $\text{EAFDL}^{\text{clus}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\lambda\,\text{MTTDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;&\approx\; \sqrt{\frac{2\pi hx}{(1-h)n}}\\ &\left[\left(\frac{h\min\left(\frac{\phi}{1-h},1\right)b}{\lambda\,c}\right)^{h}(1-h)^{1-h}\right]^{xn}\,\frac{[E(X)]^{hxn}}{E(X^{hxn})}\;,\end{aligned} \tag{76}$$

$$\begin{aligned}\text{EAFDL}^{\text{clus}}_{\text{approx}}(B_{\max})/\lambda \;&\approx\; \sqrt{\frac{1-h}{2\pi hxn}}\\ &\left[\left(\frac{h\min\left(\frac{\phi}{1-h},1\right)b}{\lambda\,c}\right)^{h}(1-h)^{1-h}\right]^{-xn}\,\frac{E(X^{hxn})}{[E(X)]^{hxn}}\;,\end{aligned} \tag{77}$$

where

$$x = \frac{m}{n}\;, \tag{78}$$

$B_{\max}$ is expressed via ϕ given by (65), and h is given by (49).

Proof: It follows from (28) that

$$\text{MTTDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;=\; \text{MTTDL}^{\text{clus}}_{\text{approx}}(\infty)\;\theta^{\,\text{clus}}\;, \tag{79}$$

and

$$\text{EAFDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;=\; \frac{\text{EAFDL}^{\text{clus}}_{\text{approx}}(\infty)}{\theta^{\,\text{clus}}}\;. \tag{80}$$

It follows from (67), and using (3), (49), and (78) that

$$\theta^{\,\text{clus}} \;=\; \min\left(\frac{\phi}{1-h},1\right)^{hxn}\;. \tag{81}$$

Substituting the analytic expression obtained in [25, Equation (71)] for the term $\lambda\,\text{MTTDL}^{\text{clus}}_{\text{approx}}(\infty)$, and (81) into (79) yields (76). Subsequently, substituting (75) and (76) into (15) and using (49) and (78) yields (77). ■

C. Declustered Placement

The declustered placement scheme is a special case of a symmetric placement scheme in which k is equal to n. Consequently, for $k = n$, (39), (40), and (41) yield

$$\begin{aligned}\text{MTTDL}^{\text{declus}}(B_{\max}) \;\approx\; &\frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)!\\ &\frac{[E(X)]^{m-l}}{E(X^{m-l})}\prod_{u=1}^{m-l}\left(\frac{n-u}{m-u}\right)^{m-l-u}\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{n}},1\right)\;,\end{aligned} \tag{82}$$

$$\text{EAFDL}^{\text{declus}}(B_{\max}) \approx \lambda \left[\frac{(l+1)\,\lambda\, c}{b}\right]^{m-l} \frac{1}{(m-l)!} \frac{E(X^{m-l})}{[E(X)]^{m-l}} \prod_{u=1}^{m-l} \left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Bigg/ \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right), \tag{83}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and

$$E(H)^{\text{declus}} \approx \left(\frac{l}{m} \prod_{u=1}^{m-l} \frac{m-u}{n-u}\right) c \tag{84}$$

$$= \frac{l\,(m-1)!\,(n-m+l-1)!}{m\,(n-1)!\,(l-1)!}\, c\,. \tag{85}$$

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 7: For large values of n, m, l, and $m-l$, $\text{MTTDL}^{\text{declus}}$ normalized to $1/\lambda$ can be approximated as follows:

$$\begin{aligned}\log\left(\lambda\,\text{MTTDL}^{\text{declus}}_{\text{approx}}(B_{\max})\right) \approx\ & + n^2\,\frac{W(h,x)}{2} + n\left\{hx\,\log\left(\frac{hx\sqrt{x}\,n\,b}{e\,[(1-h)x\,n+1]\,\lambda\,c}\right)\right. \\ & \left. + \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right\} \\ & - \frac{1}{8}\left[h(1-x) - \log\left(\frac{1-h}{1-hx}\right)\right] + \log\left(\sqrt{\frac{2\pi hx}{n}}\right) \\ & - \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{[E(X)]^{hxn}}{E(X^{hxn})}\right)\,,\end{aligned} \tag{86}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (78), (52), and (48), respectively.

Proof: Immediate from Proposition 3 by replacing k with n and using (78). ■

Proposition 8: For large values of n, m, l, and $m-l$, the $\text{EAFDL}^{\text{declus}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\log\left(\text{EAFDL}^{\text{declus}}_{\text{approx}}(B_{\max})/\lambda\right) \approx\ & - n^2\,\frac{W(h,x)}{2} \\ & + n\left\{hx\,\log\left(\frac{e\,[(1-h)x\,n+1]\,\lambda\,c}{h\,\sqrt{x}\,n\,b}\right) + \log\left(\frac{(1-hx)^{1-hx}}{(1-h)^{(1-h)x}}\right)\right. \\ & \left. - \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) - \widehat{\phi}\right\} \\ & + \frac{1}{8}\,h(1-x) + \log\left(\sqrt{\frac{1}{2\pi hxn}}\left(\frac{1-h}{1-hx}\right)^{\frac{3}{8}}\right) \\ & + \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{E(X^{hxn})}{[E(X)]^{hxn}}\right)\,,\end{aligned} \tag{87}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (78), (52), and (48), respectively.

Proof: Immediate from Proposition 4 by replacing k with n and using (78). ■

Proposition 9: For large values of n, m, l, and $m-l$, $E(H)^{\text{declus}}$ normalized to c can be approximated as follows:

$$\log\left(E(H)^{\text{declus}}_{\text{approx}}/c\right) \approx \log\left((1-h)\sqrt{\frac{1-h}{1-hx}}\right) + n\,V(h,x)\,, \tag{88}$$

where h, x, and $V(h,x)$ are given by (49), (78), and (60), respectively.

Proof: Immediate from Proposition 5 by replacing k with n and using (78). ■

IV. Reliability Optimization

For given l, m, n, and ϕ, we identify the placement scheme that offers the best reliability in terms of the MTTDL, EAFDL, and $E(H)$ metrics. In Section III, we identified the optimal placement scheme within the class of symmetric placement schemes when $m < k \leq n$. The corresponding reliability achieved should be compared with the one achieved by the clustered placement scheme when $k = m < n$. For the comparison to be meaningful, there should be at least two clustered groups, which implies that $m \leq n/2$, or, by also using (3) and (4),

$$1 \leq l < m \qquad \text{and} \qquad 1 \leq m-l < m \leq \frac{n}{2}\,. \tag{89}$$

A. *Maximizing* MTTDL

To obtain the optimal MTTDL value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then the optimal MTTDL value is equal to the $\text{MTTDL}^{\text{sym}}_{\hat{k_s}}$ value obtained by a symmetric placement where $k = \hat{k_s}$. If m divides n, then we need to compare the $\text{MTTDL}^{\text{sym}}_{\hat{k_s}}$ value with the $\text{MTTDL}^{\text{clus}}$ value obtained by the clustered placement with $k = m$. From (39) and (68), it follows that the ratio $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ of these two values is given by

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\triangleq \frac{\text{MTTDL}^{\text{sym}}_{\hat{k_s}}}{\text{MTTDL}^{\text{clus}}} \\ &\approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \\ &\quad \prod_{u=1}^{m-l} \left(\frac{\hat{k_s}-u}{m-u}\right)^{m-l-u} \min\left(\frac{\phi}{1-\frac{u}{\hat{k_s}}}, 1\right). \end{aligned} \tag{90}$$

Remark 8: It follows from (90) that the placement that maximizes MTTDL is not dependent on λ, c, or the rebuild time distribution of X, but is dependent on b and $B_{\max}$ only through ϕ, that is, the ratio $B_{\max}/b$.

The optimal placement is given by the following proposition.

Proposition 10: For given l, m, n, and ϕ, and for any λ, c, b, and rebuild time distribution of X, the value of k ($m \leq$

$k \leq n$), denoted by $\hat{k}$, that maximizes MTTDL_k is given by

$$\hat{k} = \begin{cases} m\,, & \text{for } m-l=1 \text{ and } n = j\,m\,, \text{ for some } j \in \mathbb{N} \\ \hat{k_s}\,, & \text{for } m-l=1 \text{ and } n \neq j\,m\,, \text{ for all } j \in \mathbb{N} \\ m\,, & \text{for } l=1,\ m=3,\ n=3\,j \text{ with } 2 \leq j \leq 11\,, \\ & \qquad \text{and } \phi < \frac{2\sqrt{n-2}}{n} \\ m\,, & \text{for } l=2,\ m=4,\ n=8,\ \phi < \frac{3\sqrt{3}}{8} = 0.649 \\ m\,, & \text{for } l=2,\ m=4,\ n=12,\ \phi < \frac{\sqrt{5}}{4} = 0.559 \\ m\,, & \text{for } l=3,\ m=5,\ n=10,\ \phi < \frac{4\sqrt{2}}{5\sqrt{3}} = 0.653 \\ m\,, & \text{for } l=1,\ m=4,\ n=8,\ \phi < \frac{1}{4}\sqrt[3]{\frac{15}{7}} = 0.322 \\ n\,, & \text{otherwise}\,, \end{cases} \tag{91}$$

where $\hat{k_s}$ is given by (43).

Proof: See Appendix C. ■

B. Minimizing EAFDL

To obtain the optimal EAFDL value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then, according to Proposition 2, the optimal EAFDL value is obtained by the declustered placement ($k = n$). If m divides n, then we need to compare the $\text{EAFDL}^{\text{declus}}$ value with the $\text{EAFDL}^{\text{clus}}$ value obtained by the clustered placement with $k = m$. From (83) and (74), it follows that the ratio $r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}}$ of these two values is given by

$$\begin{aligned} r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} &\triangleq \frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \\ &\approx (l+1)^{m-l}\ \frac{(l-1)!}{(m-1)!}\ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \\ &\quad \prod_{u=1}^{m-l} \left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Bigg/ \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right). \end{aligned} \tag{92}$$

Remark 9: It follows from (92) that the placement that minimizes EAFDL is not dependent on λ or c. Moreover, the ratio $r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}}$ is dependent on b and B_{max} only through ϕ, that is, the ratio B_{max}/b.

The optimal placement is given by the following proposition.

Proposition 11: For any l, m, n ($n > m$), ϕ, λ, c, b, and rebuild time distribution of X, the value of k ($m \leq k \leq n$) that minimizes the EAFDL_k is equal to n, which corresponds to the declustered placement scheme.

Proof: See Appendix D. ■

C. Minimizing $E(H)$

To obtain the optimal $E(H)$ value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then, according to Proposition 2, the optimal $E(H)$ value is obtained by the declustered placement ($k = n$). If m divides n, then we need to compare the $E(H)^{\text{declus}}$ value with the $E(H)^{\text{clus}}$ value obtained by the clustered placement with $k = m$.

From (75) and (84) and using (89), it follows that

$$r^{\text{declus,H}}_{\text{clus,H}} \triangleq \frac{E(H)^{\text{declus}}}{E(H)^{\text{clus}}} \approx \prod_{u=1}^{m-l} \frac{m-u}{n-u} < 1\,. \tag{93}$$

Remark 10: It follows from (92) that the placement that minimizes $E(H)$ is not dependent on λ, c, b, B_{max} (or, consequently, ϕ), nor on the rebuild time distribution of X.

The optimal placement is given by the following proposition.

Proposition 12: For any l, m, n ($n > m$), ϕ, λ, c, b, and rebuild time distribution of X, the value of k ($m \leq k \leq n$) that minimizes the $E(H)_k$ is equal to n, which corresponds to the declustered placement scheme.

Proof: Immediate from (93). ■

V. Optimal System Configuration

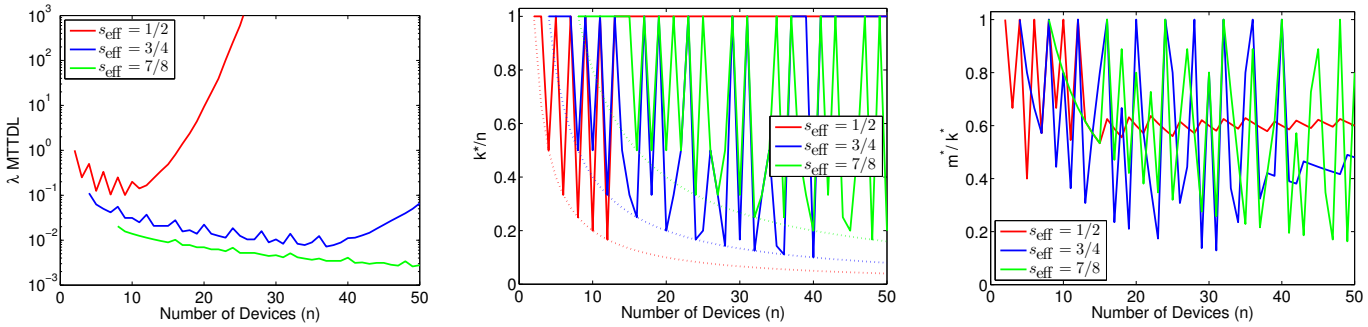
We address the issue of maximizing the reliability of a system storing an amount U of user data under a given storage efficiency s_{eff} and bandwidth constraint factor ϕ. The required number of devices n is then determined by (2). Consequently, the parameters to be specified are l, m, and k. However, these parameters are dependent. More specifically, according to (1), $l = s_{\text{eff}}\,m$. Also, given l and m, the optimal value $\hat{k}$ of k was obtained in Section IV. Consequently, to maximize system reliability, it suffices to determine the appropriate value m^* of m for the optimal codeword length. Then the optimal value k^* for the parameter k is obtained by Propositions 10, 11, and 12. Next, using a specific example, we will show that for MTTDL, we may find that $m^* \leq k^* < n$, which implies that optimality may be achieved by multiple groups, whereas for EAFDL and $E(H)$, optimality is always achieved by a single group as expressed by the following proposition.

Proposition 13: For any n, λ, c, b, B_{max} (and, consequently, ϕ), and rebuild time distribution of X, the optimal value k^* for the parameter k that minimizes the EAFDL or the $E(H)$ is equal to n.

Proof: Let us first consider the EAFDL metric, where m^* and k^* are the values that minimize it. We consider the following two cases for m^*. If $m^* < n$, then, invoking Proposition 11 with $m = m^*$, the value of k ($m^* \leq k \leq n$) that minimizes the EAFDL_k is equal to n, which implies that $k^* = n$. If $m^* = n$, then, owing to (5), it follows that $m^* = k^* = n$. Similarly, from Proposition 12, it follows that the optimal value k^* for the parameter k that minimizes the $E(H)$ is equal to n. ■

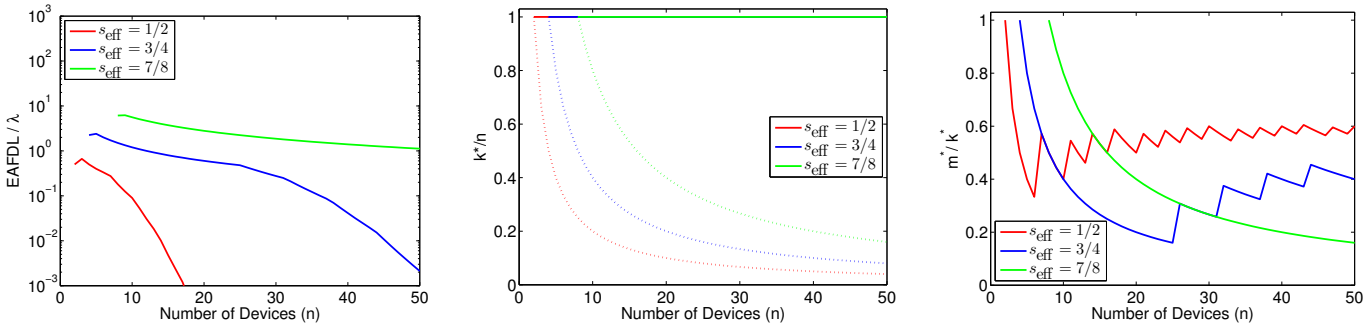
Consequently, for EAFDL and $E(H)$, the optimal placement is always the clustered or declustered one, whereas for MTTDL it may also be the symmetric one.

An alternative way to determine the optimal values m^* and k^* for the parameters m and k, respectively, is first to determine the optimal codeword length m^*_k for any given k. Note that from (39), (40), and (41), it follows that m^*_k depends on k, but not on the storage system size n. Subsequently, the optimal value of k^* can be determined by considering all possible values for k, along with the corresponding values m^*_k, and identifying the pair (k, m^*_k) that optimizes the reliability metric.



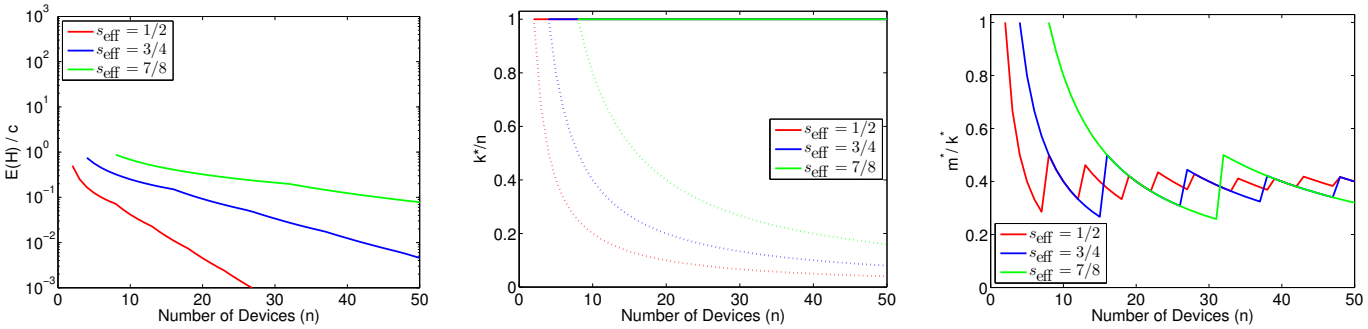
(a) Optimal MTTDL (b) k^*/n (c) m^*/k^*

Figure 4. Optimization of MTTDL vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$, $\phi = 0.001$ and deterministic rebuild times.



(a) Optimal EAFDL (b) k^*/n (c) m^*/k^*

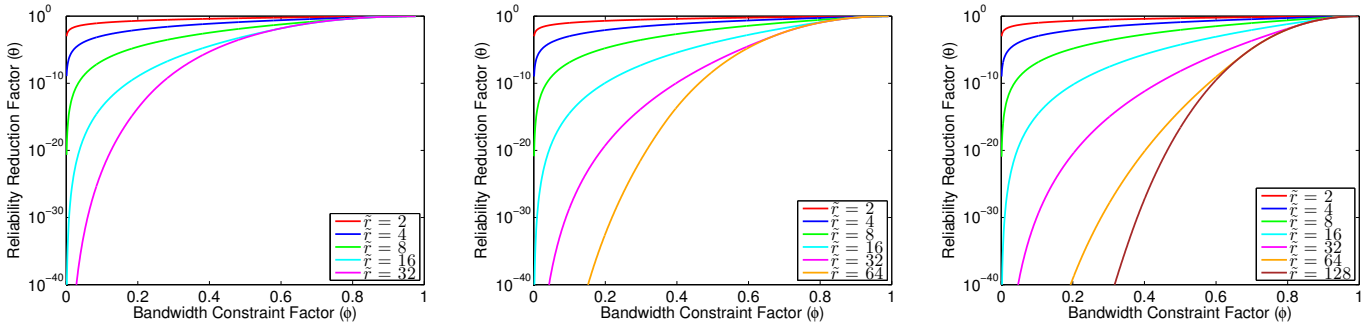
Figure 5. Optimization of EAFDL vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$, $\phi = 0.001$ and deterministic rebuild times.



(a) Optimal $E(H)$ (b) k^*/n (c) m^*/k^*

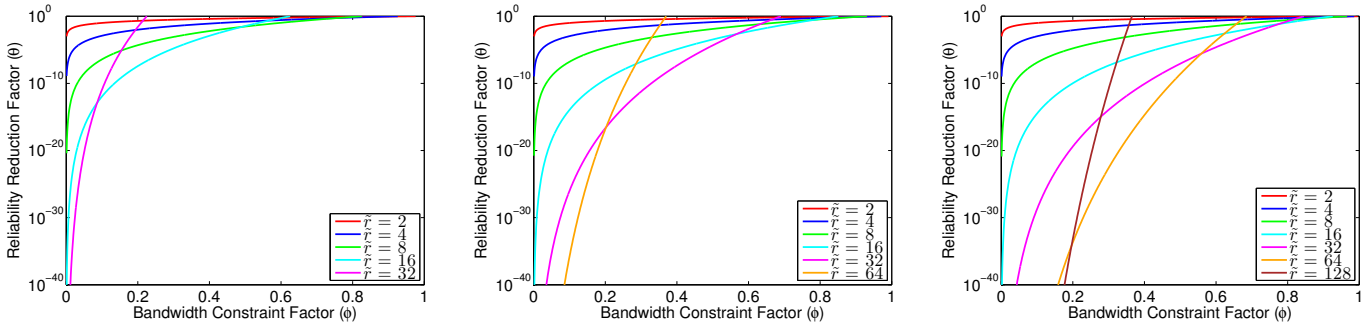
Figure 6. Optimization of $E(H)$ vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$ and $\phi = 0.001$.

Next, we consider a storage system for which it holds that $\lambda/\mu = \lambda c/b = 0.001$ and $\phi = 0.001$. We identify the optimal group sizes k^* and the optimal codeword lengths m^* that optimize the MTTDL metric for various system sizes, assuming that the rebuild time distribution is deterministic. The optimal normalized λMTTDL, EAFDL$/\lambda$, and $E(H)/c$ values along with the corresponding normalized values k^*/n and m^*/k^* are shown in Figures 4, 5, and 6, respectively, as a function of system size. From Figure 4(a) we observe that, for a given storage efficiency s_{eff}, MTTDL initially decreases and then increases as n increases. For $s_{\text{eff}} = 7/8$, MTTDL starts increasing when $n \geq 115$, which is not shown in the figure. Figure 4(b) shows the ratio of k^* to n. Given that $k^* \leq n$, the maximum value of this ratio is equal to 1. Also, k^* cannot be less than the minimum codeword length, which is equal to 2, 4 and 8, for $s_{\text{eff}} = 1/2, 3/4$ and $7/8$, respectively. Therefore, k^*/n cannot be less than $2/n$, $4/n$ and $8/n$, as indicated by the dotted lines for $s_{\text{eff}} = 1/2, 3/4$ and $7/8$, respectively. Note that when a point lies on a dotted line, that is, when k^* is equal to the minimum codeword length, then the optimal codeword length m^*, which according to (5) cannot exceed k^*, is also equal to the minimum codeword length. This implies that $k^* = m^*$ and the optimal placement is the clustered one. In this case, the ratio m^*/k^* is equal to 1, as shown in Figure 4(c). For instance, for $n = 8$ and $s_{\text{eff}} = 3/4$, $k^*/n = 0.5$, that is, $k^* = m^* = 4$, and MTTDL is maximized when we consider two groups with a clustered placement within each group. However, we see in Figure 4(b) that, for $n = 10$ and $s_{\text{eff}} = 3/4$, k^*/n is still equal to 0.5, which means that the optimal group size is now equal to 5, and



(a) $k = 40$ (b) $k = 100$ (c) $k = 200$

Figure 7. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; symmetric placement.



(a) $m = 40$ (b) $m = 100$ (c) $m = 200$

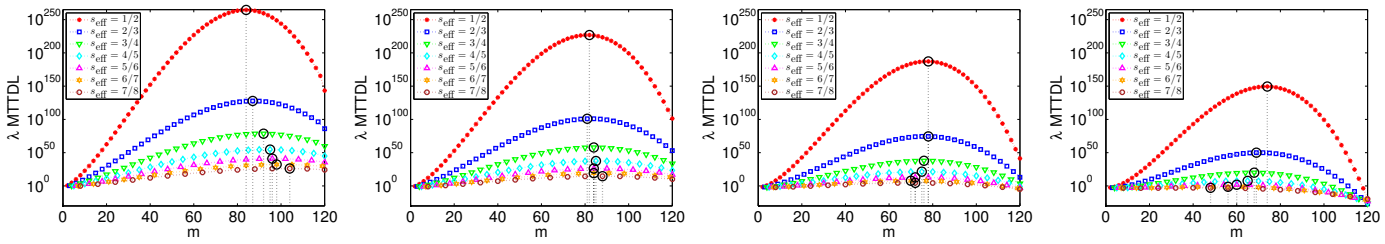
Figure 8. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; clustered placement.

the optimal codeword length remains equal to 4. In this case it holds that $m^*/k^* = 0.8$, as shown in Figure 4(c), and MTTDL is maximized when we consider two groups with a symmetric placement within each group. Note also that for $s_{\text{eff}} = 1/2$ and for system sizes that contain an even number of devices not exceeding 12, MTTDL is maximized by considering group sizes of two under a clustered placement $(k^* = m^* = 2)$. By contrast, for an odd number of devices, MTTDL is maximized by considering a single group and the declustered placement $(k^* = n)$. In particular, the optimal codeword lengths are $m^* = 2, 2, 4, 6$, and 6 for $n = 3, 5, 7, 9$, and 11, respectively. But when the number of devices exceeds 12, MTTDL is maximized by considering a single group under declustered placement and codewords whose lengths are about 60% of the system size $(k^* = n$ and $m^* \approx 0.6\, n)$. For $s_{\text{eff}} = 7/8$ and $n = 42$, it turns out that $k^* = 14$ or, equivalently, $k^*/n = 1/3$, and $m^* = 8$ or, equivalently, $m^*/k^* = 4/7$. Thus, MTTDL is maximized by considering three groups with a group size of 14 and a symmetric placement of codewords of length 8 containing seven user-data symbols each. Considering $l = 7$, $m = 8$, and $n = 42$, we confirm the optimal value of k by invoking (91), which in this case yields $\hat{k} = \hat{k_s}$, then using (43), which yields $\hat{k_s} = k_m$, and finally using (45), which yields $k_m = 14$. In general, the declustered placement is optimal, except in the cases of small n and ϕ where another placement may be optimal. However, this does not happen in the case of minimizing the EAFDL and $E(H)$ metrics. According to Proposition 13, and as shown in Figures 5(b) and 6(b), for all values of n, EAFDL and $E(H)$ are minimized by a single group $(k^* = n)$ and the clustered or declustered placement depending on whether n is equal to or exceeds the minimum codeword length, respectively.

VI. Numerical Results

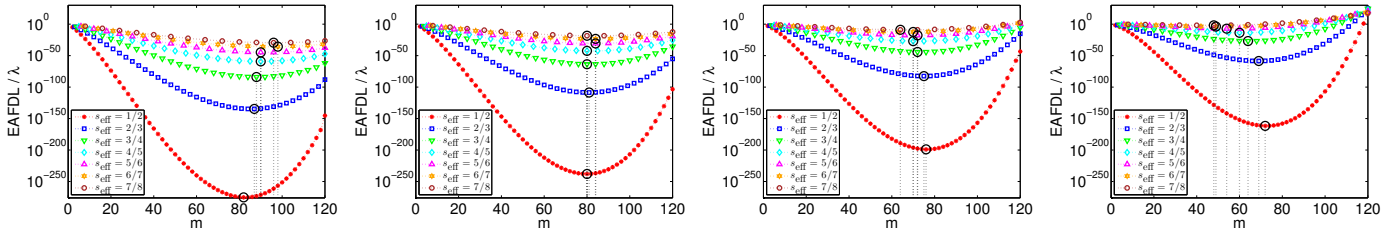
First, we assess the reduction in reliability owing to bandwidth constraints. The reliability reduction factor θ is obtained by (38) and (67) for the symmetric and clustered placements, respectively, and shown in Figures 7 and 8 as a function of the bandwidth constraint factor. For a symmetric placement scheme, Figure 7 demonstrates that as the group size k increases, the reliability reduction factor θ decreases and the magnitude of the reduction is more pronounced for larger values of $\tilde{r}$. Clearly, if codewords are spread over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process, the system reliability is affected and a drastic reliability degradation occurs as the system size increases. In contrast, according to Remark 7, the reliability of a clustered placement scheme remains unaffected for $\phi \geq l/m = (m - \tilde{r} + 1)/m$. This is due to the fact that the effective rebuild bandwidth is significantly smaller because the rebuilds are not distributed, but performed directly on a spare device. However, as Figure 8 demonstrates for $\phi < l/m$, the reliability reduction factor drops sharply, especially for large values of $\tilde{r}$.

Next, we consider a storage system of a given size and assess its reliability for various codeword configurations, storage efficiencies, and network rebuild bandwidth constraints. In particular, we consider a system containing 120 devices under a declustered placement scheme $(k = n = 120)$, which according to Remark 5 is optimal within the class of symmetric schemes. The amount U of user data stored is determined by



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 9. Normalized MTTDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

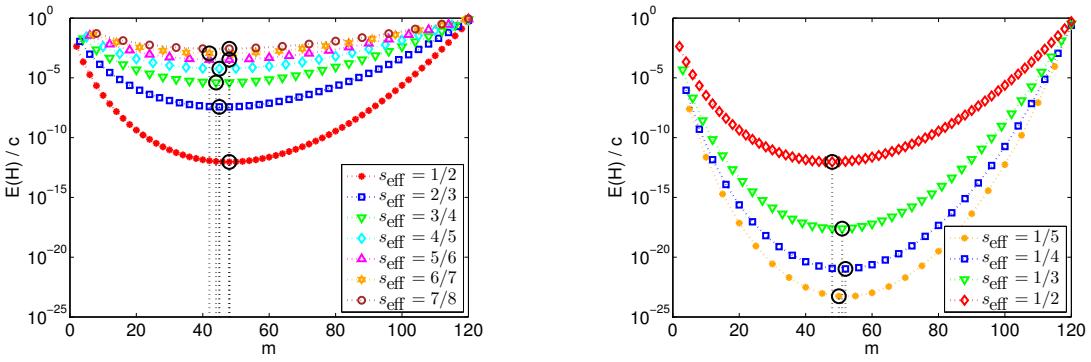
Figure 10. Normalized EAFDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$ and deterministic rebuild times.

the storage efficiency s_{eff} via (2). As discussed in Section II-E, the analytical reliability results obtained are accurate when the storage devices are highly reliable, that is, when the ratio λ/μ of the mean rebuild time $1/\mu$ to the mean time to failure of a device $1/\lambda$ is very small. We proceed by considering systems for which it holds that $\lambda/\mu = \lambda\, c/b = 0.001$ and the distribution of the rebuild time X is deterministic, that is, $E(X^{m-l}) = [E(X)]^{m-l}$.

The combined effect of the network rebuild bandwidth constraint and the storage efficiency on the normalized λ MTTDL measure is obtained by (82) and shown in Figure 9 as a function of the codeword length. In particular, when the codeword length is equal to the system size $(m = k = n)$, the placement becomes clustered and the normalized λ MTTDL measure is obtained by (68). Four cases for the network rebuild bandwidth constraint were considered: $\phi = 1$ corresponds to the case where there is no network rebuild bandwidth constraint given that $N_b \geq k = 120$ or, equivalently, $B_{\max} \geq k\, b = 120\, b$; $\phi = 0.1$, $\phi = 0.01$, and $\phi = 0.001$ correspond to the cases where $N_b = 0.1\, k = 12$, $N_b = 0.01\, k = 1.2$, and $N_b = 0.001k = 0.12$ or, equivalently, $B_{\max} = 0.1\, k\, b = 12\, b$, $B_{\max} = 0.01\, k\, b = 1.2\, b$, and $B_{\max} = 0.001\, k\, b = 0.12\, b$, respectively. The values for the storage efficiency are chosen to be fractions of the form $z/(z+1)$, $z = 1, \ldots, 7$, such that the first point of each of the corresponding curves is associated with the single-parity $(z, z+1)$-erasure code, and the second point of each of the corresponding curves is associated with the double-parity $(2z, 2z+2)$-erasure code.

For all values of ϕ considered, we observe that MTTDL increases as the storage efficiency s_{eff} decreases. This is because, for a given m, decreasing s_{eff} implies decreasing l, which in turn implies increasing the parity symbols $m - l$ and consequently improving the MTTDL. Furthermore, for a given storage efficiency s_{eff}, MTTDL decreases by orders of magnitude as the maximum permitted network rebuild bandwidth decreases. We now proceed to identify the optimal codeword length m^* that maximizes MTTDL for a given bandwidth constraint and storage efficiency. The optimal codeword length is dictated by two opposing effects on reliability. On the one hand, larger values of m imply that codewords can tolerate more device failures, but on the other hand, they result in a higher exposure degree to failure as each of the codewords is spread across a larger number of devices. In Figure 9, the optimal values m^* are indicated by the circles, and the corresponding codeword lengths are indicated by the vertical dotted lines. By comparing Figures 9(a), (b), (c), and (d), we deduce that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the maximum MTTDL value of $4{\times}10^{78}$ is obtained when $m = m^* = 92$. However, in the case of $\phi = 0.1$, the maximum MTTDL value of $6{\times}10^{57}$ is obtained for $m^* = 84$. The reason for the reduction of the optimal codeword length is that $\tilde{r}$ increases with increasing m for a given value of s_{eff}, which, according to Remark 3, results in a lower reliability reduction factor. Thus, the reliability reduction factor corresponding to $m = 92$ is lower than the one corresponding to $m = 84$, which in turn causes MTTDL for $m = 92$ to no longer be optimal as it becomes lower than the one for $m = 84$. Note that for $m = 84$ and $s_{\text{eff}} = 3/4$, it follows from (1) and (3) that $l = 63$ and $\tilde{r} - 1 = 21$. From (38), and given that $u \leq \tilde{r} - 1 = 21 \ll k = 120$, such that $\phi/(1 - u/k) \approx \phi$, it now follows that $\theta \approx \phi^{\tilde{r}-1} = 0.1^{21} = 10^{-21}$, which implies that the reliability is reduced by 21 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the maximum MTTDL values of $6{\times}10^{37}$ and $8{\times}10^{19}$ are obtained for $m^* = 76$ and $m^* = 68$, respectively.

The combined effect of the network rebuild bandwidth constraint and the storage efficiency on the normalized $\text{EAFDL}^{\text{declus}}/\lambda$ measure is obtained by (74) and (83), and



(a) $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$

(b) $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$

Figure 11. Normalized $E(H)$ vs. codeword length; $n = k = 120$.

shown in Figure 10 as a function of the codeword length. We observe that EAFDL increases as the storage efficiency s_{eff} decreases. Furthermore, for a given storage efficiency s_{eff}, EAFDL increases by orders of magnitude as the maximum permitted network rebuild bandwidth decreases. In fact, for $\phi = 0.01$ and $\phi = 0.001$, Figures 10(c) and (d) show that the EAFDL can be greater than 1.

Remark 11: Although the fraction of data loss never exceeds 1, EAFDL can exceed 1 because it expresses the annual fraction of data loss, which also takes into account the frequency of data losses.

Similarly to the case of MTTDL, by comparing Figures 10(a), (b), (c), and (d), we observe that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the minimum EAFDL value of 10^{-84} is obtained when $m = m^* = 88$. However, in the case of $\phi = 0.1$, the minimum EAFDL value of $2{\times}10^{-64}$ is obtained for $m^* = 80$, which implies that the reliability is reduced by 20 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the minimum EAFDL values of $7{\times}10^{-45}$ and 10^{-27} are obtained for $m^* = 72$ and $m^* = 64$, respectively. By comparing Figures 9 and 10, we deduce that in general the optimal codeword lengths m^*_{MTTDL} (for MTTDL) and m^*_{EAFDL} (for EAFDL) are similar.

The effect of the storage efficiency on the normalized $E(H)/c$ measure is obtained by (75) and (84), and shown in Figure 11 as a function of the codeword length. Note that according to Remark 1, neither the network rebuild bandwidth constraint nor the rebuild time distribution affects this metric. We observe that $E(H)$ increases as the storage efficiency s_{eff} increases.

Remark 12: From (27), and recalling (1), (5) and the fact that V_u is a fraction, we deduce that $E(H)/c \leq s_{\text{eff}} < 1$, which implies that $E(H) < c$.

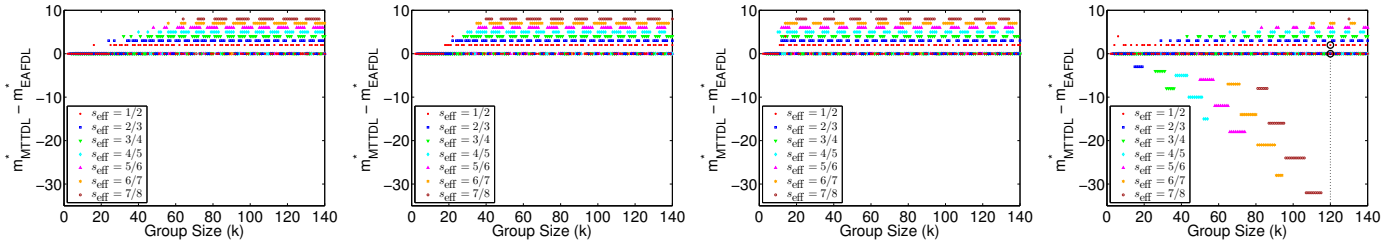
Reducing $B_{\max}$ or, equivalently, ϕ affects the optimal codeword lengths for MTTDL and EAFDL as follows.

Proposition 14: For given n, k, and s_{eff}, and for the MTTDL and EAFDL reliability metrics, the optimal codeword length m^* decreases with decreasing ϕ.

Proof: Consider two bandwidth constraint factors ϕ_1 and ϕ_2 with $\phi_1 > \phi_2$. Let m_1^* and m_2^* be the corresponding optimal codeword lengths for the MTTDL metric. We shall now show that $m_1^* \geq m_2^*$.

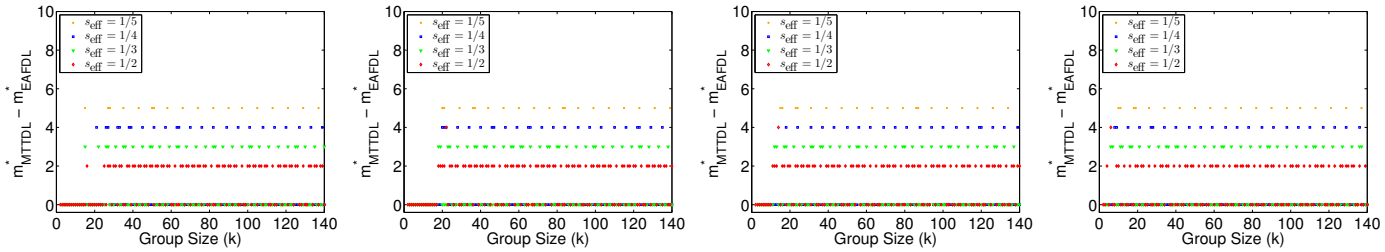
As m_1^* is the optimal codeword length for ϕ_1, it holds that $\text{MTTDL}(\phi_1, m) \leq \text{MTTDL}(\phi_1, m_1^*)$ for all $m \geq m_1^*$. Also, from (1) and (3), it holds that $\tilde{r} = (1 - s_{\text{eff}})\, m + 1$, which implies that as m increases, so does $\tilde{r}$. From (29), it follows that $\theta^{(2)}/\theta^{(1)} = \prod_{u=1}^{\tilde{r}-1} \frac{b_u(\phi_2)}{b_u(\phi_1)}$, which, owing to the fact that $b_u(\phi_2) \leq b_u(\phi_1)\ \forall\, u$, decreases with increasing $\tilde{r}$ or, equivalently, m. Consequently, $\theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)}$ for all $m \geq m_1^*$. Also, it follows from (28) that $\text{MTTDL}(\phi_2,m)/\text{MTTDL}(\phi_1,m) = \theta_m^{(2)}/\theta_m^{(1)}$ for all values of m. From the preceding, it follows that $\text{MTTDL}(\phi_2,m)/\text{MTTDL}(\phi_1,m) = \theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)} = \text{MTTDL}(\phi_2,m_1^*)/\text{MTTDL}(\phi_1,m_1^*) \leq \text{MTTDL}(\phi_2,m_1^*)/\text{MTTDL}(\phi_1,m)$ for all $m \geq m_1^*$. Thus, $\text{MTTDL}(\phi_2,m) \leq \text{MTTDL}(\phi_2,m_1^*)$ for all $m \geq m_1^*$, which in turn implies that $m_2^* \leq m_1^*$. The proof for EAFDL is similar to that for MTTDL and is therefore omitted. ■

Figures 12 and 13 show the difference between the optimal codeword lengths for MTTDL and EAFDL. They demonstrate that the optimal codeword length for MTTDL is generally greater than or equal to that for EAFDL, with the difference being equal either to $z + 1$, the denominator of the storage efficiency fraction, or to 0. This implies that the optimal codeword lengths m^*_{EAFDL} for EAFDL are either equal to or slightly smaller than and adjacent to the optimal codeword lengths m^*_{MTTDL} for MTTDL. However, for small values of ϕ, such as $\phi = 0.001$, m^*_{MTTDL} can be smaller than m^*_{EAFDL}, as observed in Figure 12(d). This occurs only for certain group sizes that are smaller than 120, whereas for $k \geq 120$, the optimal codeword lengths follow the general trend discussed above. For example, in the case of $k = 120$, $\phi = 0.001$, and $s_{\text{eff}} = 1/2$, Figure 9(d) shows that the maximum value of MTTDL is achieved when the codeword length m is equal to 74, which implies that $m^*_{\text{MTTDL}} = 74$. Also, Figure 10(d) shows that the minimum value of EAFDL is achieved when the codeword length m is equal to 72, which implies that $m^*_{\text{EAFDL}} = 72$. The value of 72 is adjacent to 74 because when $s_{\text{eff}} = 1/2$, m cannot be equal to 73. Consequently, the difference of the optimal codeword lengths for EAFDL and MTTDL is given by 74 – 72 = 2, indicated by a circle in Figure 12(d). Similarly, for $k = 120$ and $s_{\text{eff}} = 2/3$, Figures 9(d) and 10(d) show that both the optimal MTTDL and the



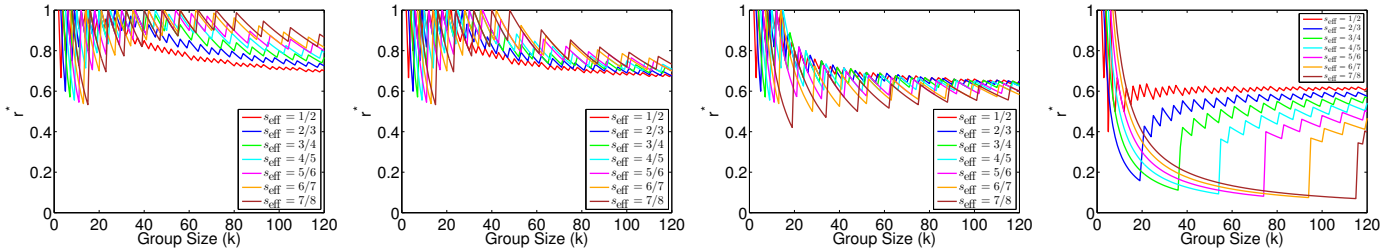
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 12. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for various storage efficiencies; $\lambda/\mu = 0.001$ and deterministic rebuild times.



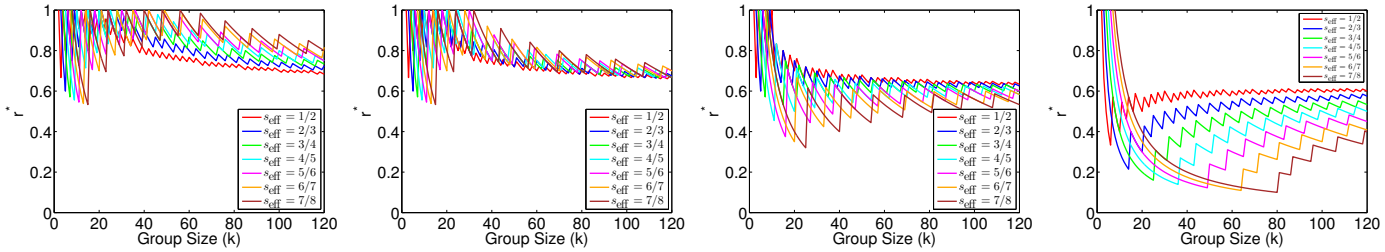
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 13. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 14. r^* for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

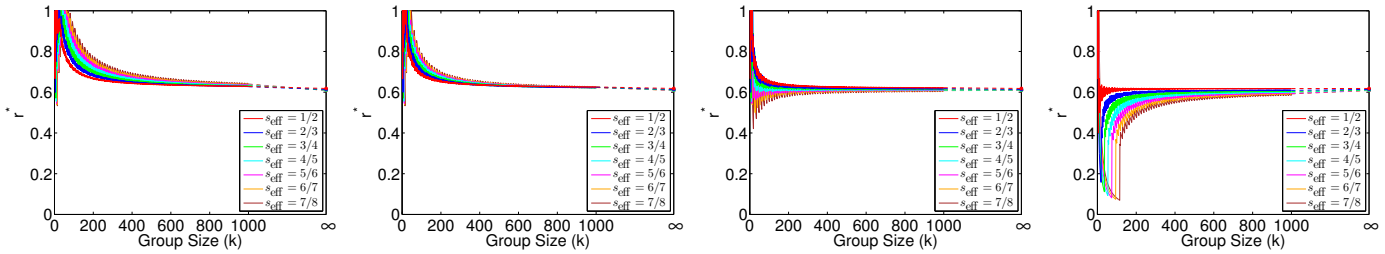
Figure 15. r^* for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.

optimal EAFDL are obtained when the codeword length is equal to 69, that is, $m^*_{\text{MTTDL}} = m^*_{\text{EAFDL}} = 69$. In this case, the difference of the optimal codeword lengths for EAFDL and MTTDL is equal to 0 , indicated by a circle in Figure 12(d).

To investigate the behavior of the optimal codeword length m^*_k with increasing group size k, we proceed by considering the normalized optimal codeword length r^*, namely, the ratio of m^*_k to k:

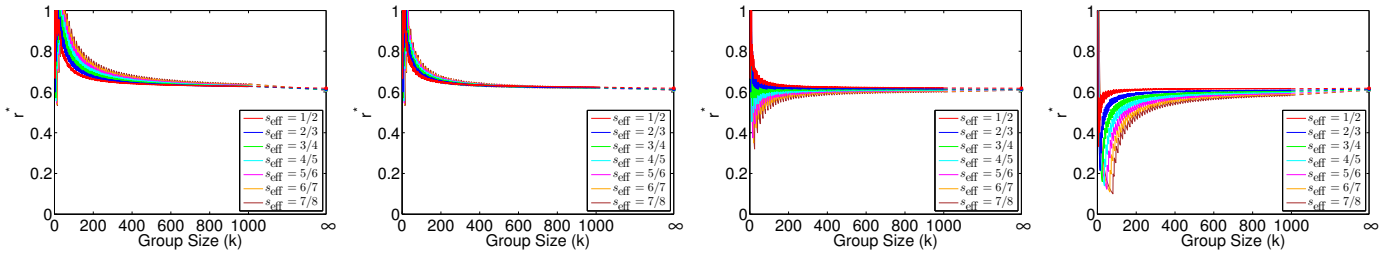
$$r^* \triangleq \frac{m^*_k}{k} . \tag{94}$$

The r^* values for the MTTDL and EAFDL metrics are shown in Figures 14 and 15, respectively, for various storage efficiencies and network rebuild bandwidth constraints. According to Proposition 14, for any storage efficiency s_{eff} and for any given group size k, the optimal codeword lengths and, consequently, the r^* values decrease with decreasing ϕ. Also, when the bandwidth constraint factor ϕ is small, the r^* values first decrease and then gradually increase with increasing k. The initial decrease is due to the fact that the optimal codeword length m^* remains fixed and equal to $z + 1$, which is the minimum possible codeword length for the storage efficiency



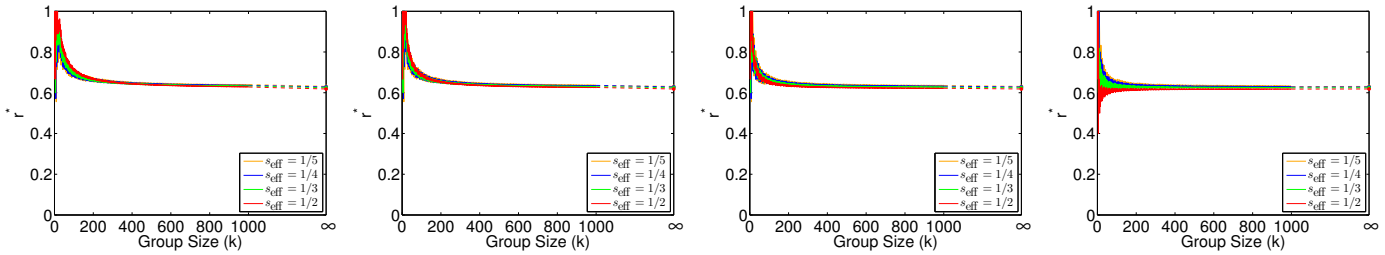
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 16. r^* for MTTDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



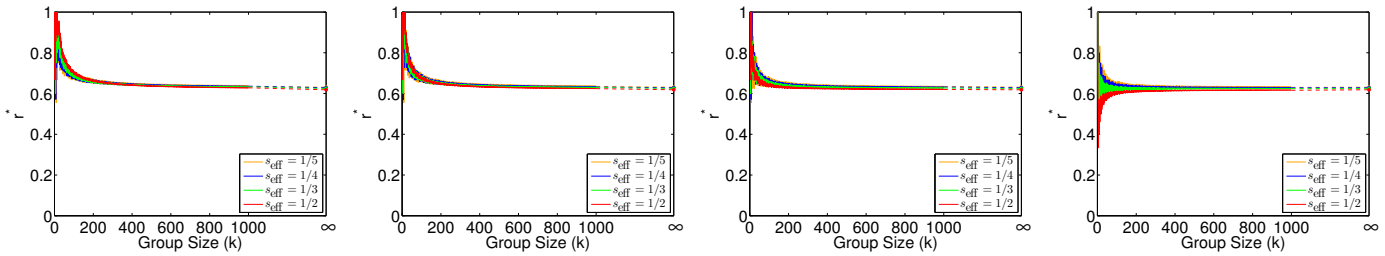
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 17. r^* for EAFDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 18. r^* for MTTDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 19. r^* for EAFDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.

fractions $z/(z+1)$, $z = 1, \ldots, 7$. For example, in the case of $s_{\text{eff}} = 7/8$ and $\phi = 0.001$, $m^* = 8$ for $k \leq 115$ in the case of MTTDL, or for $k \leq 80$ in the case of EAFDL, as shown in Figures 14(d) and 15(d), respectively.

The r^* values for the MTTDL and EAFDL metrics for various values of the storage efficiency s_{eff} and for large values of k are shown in Figures 16, 17, 18, and 19. We observe that, for a given storage efficiency and as k increases, the r^* values for MTTDL and EAFDL approach a common value, denoted by r^*_∞ and indicated by a small bullet. The r^*_∞ value is given by the following proposition.

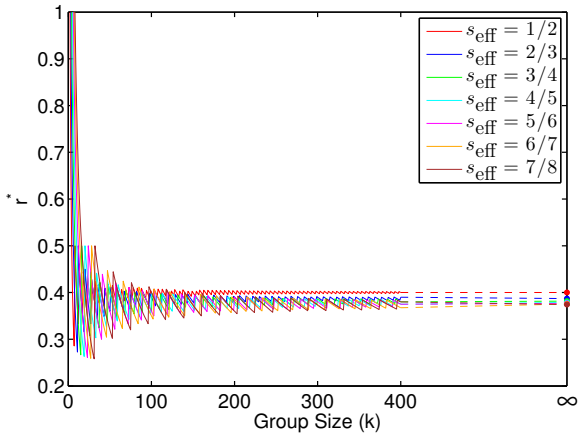
Proposition 15: As k increases, the r^* values for MTTDL and EAFDL approach r^*_∞ that satisfies the following equation:

$$Q(h, r^*_\infty) = 0 \,, \tag{95}$$

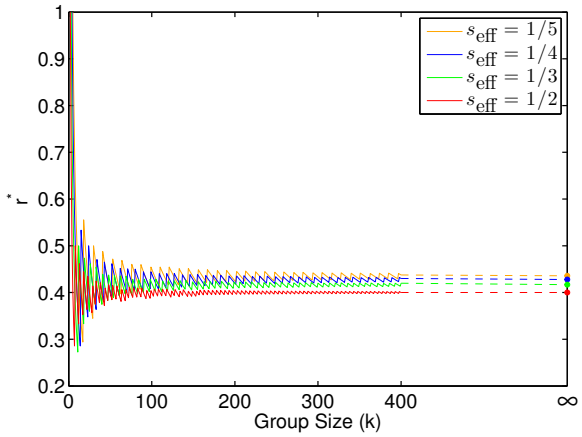
where $Q(h, x)$ is given by

$$Q(h,x) \triangleq hx + \log\left([(1-h)^{(1-h)^2} x^{h^2}]^x (1-hx)^{h(1-hx)}\right) , \tag{96}$$

and h and x are given by (49) and (50), respectively.



(a) $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5,\ 5/6, 6/7,$ and $7/8$

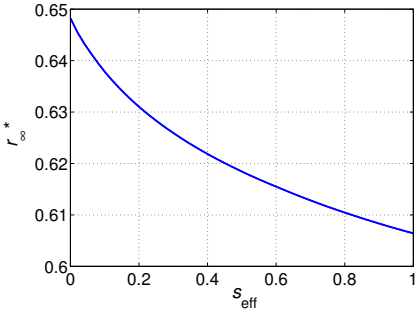


(b) $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$

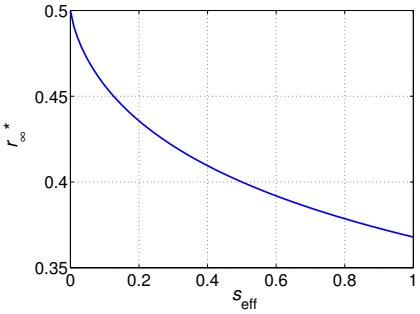
Figure 20. r^* for $E(H)$ vs. group size $k \to \infty$.

TABLE II. r^*_∞ VALUES FOR VARIOUS s_{eff}

s_{eff}			r^*_∞	
			MTTDL and EAFDL	$E(H)$
0	=	0	0.648419	0.5
10^{-4}	=	0.0001	0.648404	0.499795
10^{-3}	=	0.001	0.648265	0.498520
10^{-2}	=	0.01	0.646985	0.490770
10^{-1}	=	0.1	0.637940	0.456298
1/8	=	0.125	0.636043	0.450268
1/7	=	0.142857	0.634788	0.446383
1/6	=	0.166667	0.633224	0.441637
1/5	=	0.2	0.631212	0.435664
1/4	=	0.25	0.628500	0.427826
1/3	=	0.333333	0.624638	0.416889
1/2	=	0.5	0.618499	0.4
2/3	=	0.666667	0.613720	0.387097
3/4	=	0.75	0.611679	0.381625
4/5	=	0.8	0.610543	0.378586
5/6	=	0.833333	0.609818	0.376650
6/7	=	0.857143	0.609316	0.375307
7/8	=	0.875	0.608946	0.374322
$1 - 10^{-1}$	=	0.9	0.608440	0.372971
$1 - 10^{-2}$	=	0.99	0.606713	0.368368
$1 - 10^{-3}$	=	0.999	0.606549	0.367928
$1 - 10^{-4}$	=	0.9999	0.606532	0.367884
1	=	1	0.606531 = $1/\sqrt{e}$	0.367879 = $1/e$



(a) MTTDL and EAFDL



(b) $E(H)$

Figure 21. r^*_∞ vs. s_{eff}.

Proof: It follows from (51) that, for large values of k, $k^2\, W(h,x)/2$ is the dominating term. Thus, MTTDL is maximized when $W(h,x)$ is maximized. According to the arguments in Appendix F of [25], it therefore holds that [25, Equation (165)]

$$r^*_\infty = \arg \max_{0<x\leq 1} W(h,x)\ . \tag{97}$$

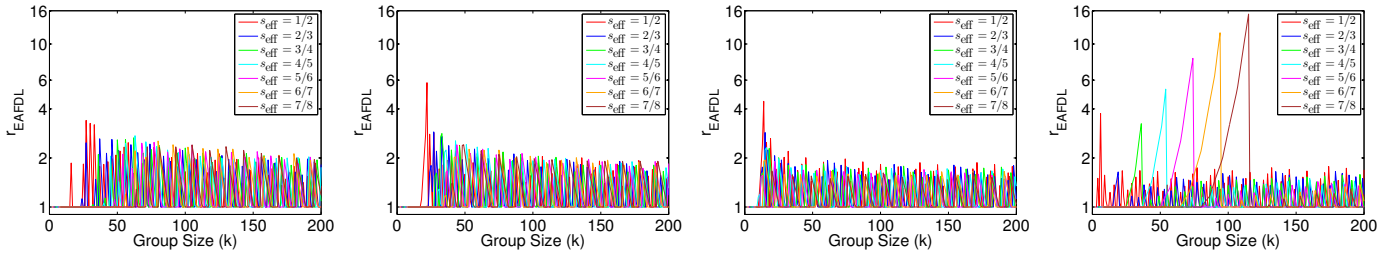
Consequently, r^*_∞ is obtained as the unique root of the equation $Q(h,x) = 0$, with respect to x, in the interval $(0,1]$, that is, [25, Equation (176)]

$$Q(h, r^*_\infty) = 0\ , \quad \text{with}\ \ r^*_\infty \in (0,1]\ , \tag{98}$$

where $Q(h,x)$ is given by (96) [25, Equation (105)]. From (55), it follows that the same rationale applies in the case of EAFDL. ■

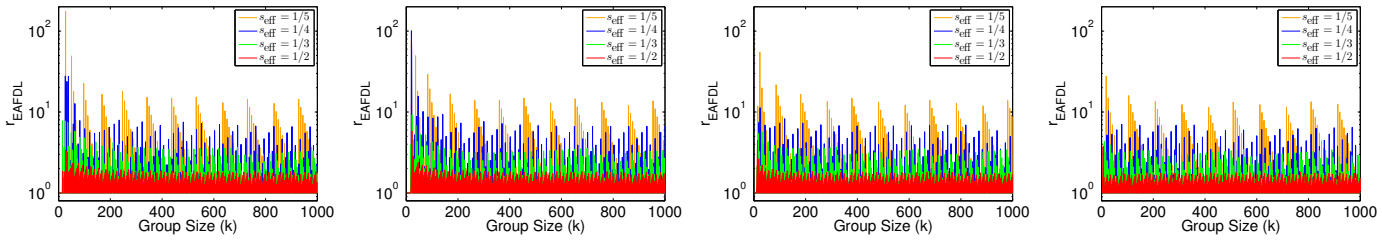
The r^* values for the $E(H)$ metric are shown in Figure 20 for various storage efficiencies and also for large group sizes. Clearly, the optimal codeword lengths for $E(H)$ are generally significantly shorter than those for MTTDL and EAFDL. We observe that, for a given storage efficiency and as k increases, the r^* values for $E(H)$ oscillate and approach a common value, denoted by r^*_∞ and indicated by a small bullet.

Remark 13: The r^*_∞ values are not affected by the bandwidth constraint factor ϕ and depend only on the storage efficiency s_{eff}. This is because the resulting reliability reduction factor θ, according to (47), is of the order $O(k)$, whereas the MTTDL and EAFDL reliability metrics, according to (51) and (55), are of the order $O(k^2)$, which is higher. Note that this also holds when the average network rebuild bandwidth is upper limited by $B_{\max}$, such that the bandwidth constraint factor ϕ is no longer constant, but, for large values of k and according



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 22. The EAFDL efficiency ratio r_{EAFDL} vs. group size; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 23. The EAFDL efficiency ratio r_{EAFDL} vs. group size; $\lambda/\mu = 0.001$ and deterministic rebuild times.

to (36), is inversely proportional to k. In this case, according to (119), the resulting reliability reduction factor θ is of the order $O(k\,\log(k))$, which is still smaller than the order $O(k^2)$. Also, according to Remark 1, the $E(H)$ metric is not affected by the bandwidth constraint factor, which implies that the r^*_∞ value for $E(H)$ is given by [25, Equation (107)]

$$r^*_\infty = \frac{1}{h + (1-h)^{-\frac{1-h}{h}}}\ , \tag{99}$$

with h given by (49).

The r^*_∞ values for the reliability metrics considered were initially derived in [25] and are included in this paper in Table II and Figure 21 for completeness. Note that the r^*_∞ values for the MTTDL and EAFDL are in the interval $[e^{-1/2} = 0.606, 0.648]$, whereas those for $E(H)$ are in the interval $[e^{-1} = 0.368, 0.5]$. Also, the r^*_∞ values decrease with increasing storage efficiency s_{eff}.

Next we examine the increase of the EAFDL metric if, instead of the optimal codeword lengths m^*_{EAFDL}, we use the codeword lengths m^*_{MTTDL} that optimize the MTTDL metric. From the preceding, it generally follows that m^*_{MTTDL} is either equal or adjacent to m^*_{EAFDL}, that is, $m^*_{\text{MTTDL}} = m^*_{\text{EAFDL}}+z+1$. We define the EAFDL *efficiency ratio* r_{EAFDL} as the ratio of $\text{EAFDL}(m^*_{\text{MTTDL}})$ to $\text{EAFDL}(m^*_{\text{EAFDL}})$, that is,

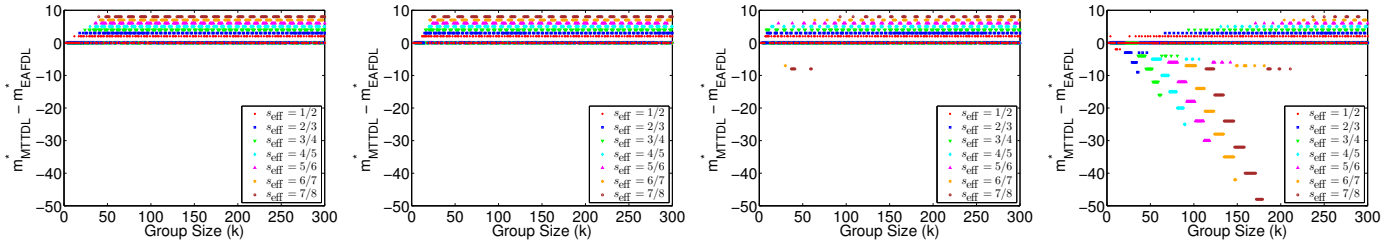
$$r_{\text{EAFDL}} \triangleq \frac{\text{EAFDL}(m^*_{\text{MTTDL}})}{\text{EAFDL}(m^*_{\text{EAFDL}})}\ , \tag{100}$$

where $\text{EAFDL}(m)$ denotes the EAFDL corresponding to a codeword length m.

The EAFDL efficiency ratios r_{EAFDL} as a function of k for various storage efficiencies and network rebuild bandwidth constraints are shown in Figures 22 and 23. We observe that for the storage efficiencies considered in Figure 22 and as k increases, the EAFDL efficiency ratios follow a periodic pattern and, for $\phi = 1$, are always less than a factor of 4. Moreover, as ϕ decreases, the EAFDL efficiency ratios tend to be less than a factor of 2, except in a few cases where they are significantly higher. Nevertheless, in all cases they are less than a factor of 16, which implies that using codewords of length m^*_{MTTDL} yields the maximum possible (optimal) MTTDL and also an EAFDL that is either optimal or of the same order. The maximum value is shown in Figure 22(d) obtained when $\phi = 0.001$, $s_{\text{eff}} = 7/8$, and $k = 115$. In this case, it holds that $m^*_{\text{MTTDL}} = 8$ and $m^*_{\text{EAFDL}} = 40$, such that $\text{EAFDL}(m^*_{\text{EAFDL}})/\lambda = \text{EAFDL}(40)/\lambda = 0.032$ and $\text{EAFDL}(m^*_{\text{MTTDL}})/\lambda = \text{EAFDL}(8)/\lambda = 0.487$, which in turn yields an EAFDL efficiency ratio r_{EAFDL} of 0.487 / 0.032 = 15.2. Also, as the storage efficiency decreases, the EAFDL efficiency ratio r_{EAFDL} increases, as shown in Figure 23. For any given storage efficiency and bandwidth constraint factor, r_{EAFDL} follows a periodic pattern and for $s_{\text{eff}} > 1/4 = 0.25$, r_{EAFDL} tends to be less than a factor of 10. Consequently, using codewords of length m^*_{MTTDL} yields an EAFDL that is of the same order of magnitude as the optimal one.

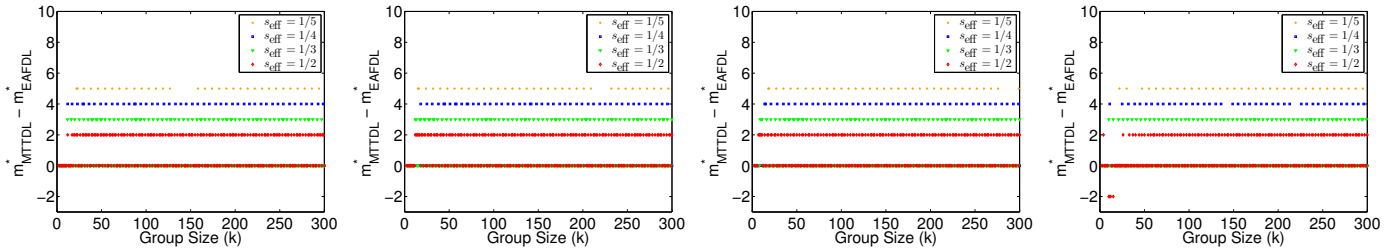
Next, we consider a system where the distribution of the rebuild time X is exponential, for which it holds that $E(X^{m-l}) = (m-l)!\,[E(X)]^{m-l}$. The combined effect of the network rebuild bandwidth constraint, the storage efficiency, and the codeword length on the reliability measures considered is similar to the case of deterministic rebuild times. Furthermore, similar to the case of deterministic rebuild times, the optimal codeword lengths m^*_{EAFDL} for EAFDL are generally either equal to or slightly shorter than and adjacent to the optimal codeword lengths m^*_{MTTDL} for MTTDL, as demonstrated in Figures 24 and 25.

The r^* values for the MTTDL and EAFDL metrics are shown in Figures 26 and 27, respectively, for various storage efficiencies and network rebuild bandwidth constraints. According to Remark 13 and Remark 13 of Appendix F of [25], as k increases, and for any storage efficiency and bandwidth constraint factor, the r^* values for MTTDL and



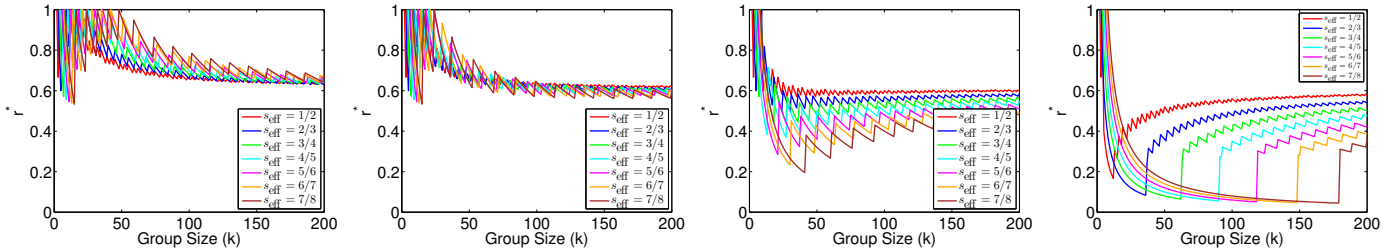
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 24. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for various storage efficiencies; $\lambda/\mu = 0.001$ and exponential rebuild times.



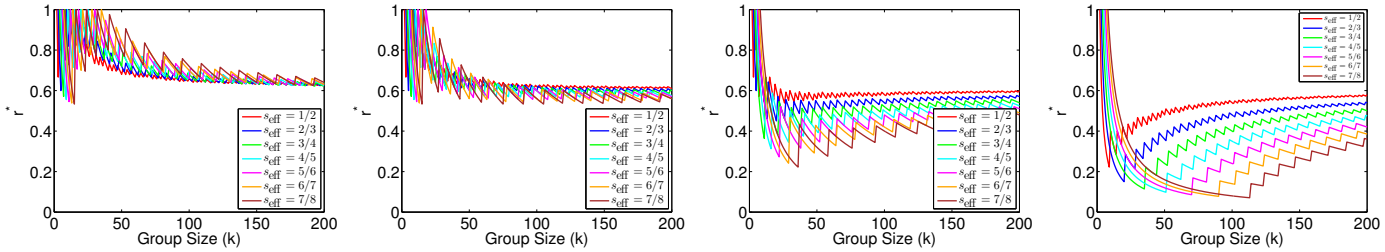
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 25. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and exponential rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 26. r^* for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.

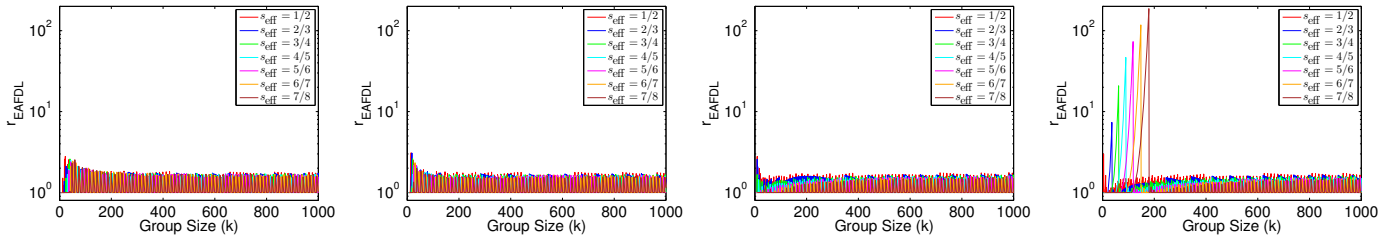


(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 27. r^* for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.

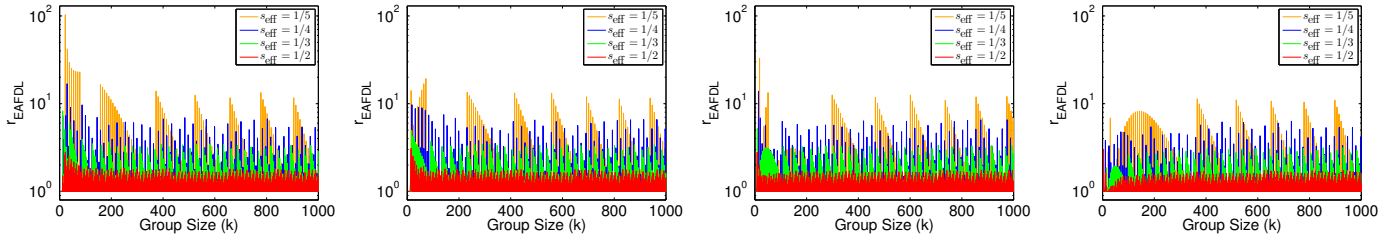
EAFDL approach a common value that is the same as the r^*_∞ value obtained in the case of deterministic rebuild times, which only depends on s_{eff} and is listed in Table II.

The EAFDL efficiency ratios r_{EAFDL} are shown in Figures 28 and 29 as a function of k for various storage efficiencies and network rebuild bandwidth constraints. We observe that as k increases, the EAFDL efficiency ratios follow a periodic pattern, as in the case of deterministic rebuild times. In particular, for the storage efficiencies considered in Figure 28, the EAFDL efficiency ratios tend to be less than a factor of 3, except in a few cases where they are significantly higher. The maximum value is shown in Figure 28(d) obtained when $\phi = 0.001$, $s_{\text{eff}} = 7/8$, and $k = 179$. In this case, it holds that $m^*_{\text{MTTDL}} = 8$ and $m^*_{\text{EAFDL}} = 56$, such that $\text{EAFDL}(m^*_{\text{EAFDL}})/\lambda = \text{EAFDL}(56)/\lambda = 0.00167$ and $\text{EAFDL}(m^*_{\text{MTTDL}})/\lambda = \text{EAFDL}(8)/\lambda = 0.31285$, which in turn yields an EAFDL efficiency ratio r_{EAFDL} of 0.31285 / 0.00167 = 187. Also, as the storage efficiency decreases, the EAFDL efficiency ratio r_{EAFDL} increases, as shown in Figure 29. Nevertheless, for $s_{\text{eff}} > 1/4 = 0.25$, r_{EAFDL} tends to be less than a factor of 10. Consequently, using codewords of length m^*_{MTTDL} yields an EAFDL that is of the same order



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 28. The EAFDL efficiency ratio r_{EAFDL} vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 29. The EAFDL efficiency ratio r_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and exponential rebuild times.

of magnitude as the optimal one.

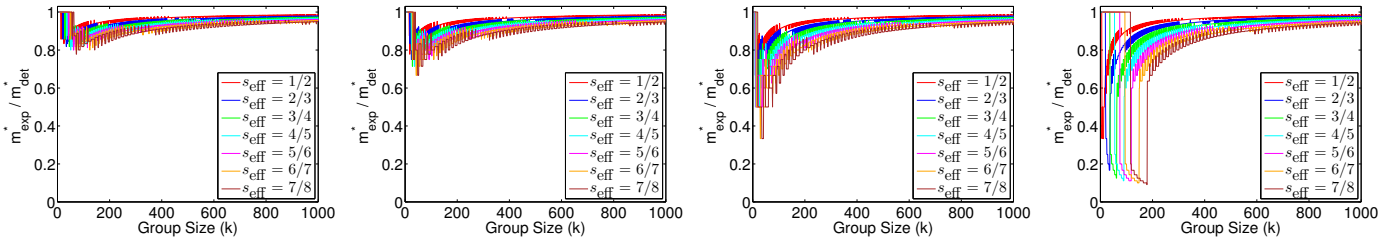
Figures 30 to 33 show the ratio of the optimal codeword length m^*_{exp} for the exponential distribution to the optimal codeword length m^*_{det} for the deterministic distribution for various storage efficiencies and network rebuild bandwidth constraints. We observe that this ratio never exceeds 1 and approaches 1 as k increases. This implies that, regardless of the rebuild bandwidth constraint, the optimal codeword length for the exponential distribution is generally smaller than the optimal codeword length for the deterministic distribution. This can be intuitively explained as follows. As previously mentioned, higher values of m result in a greater exposure degree to failure as each of the codewords is spread across a larger number of devices. The variation of exponentially distributed rebuild times results in increased vulnerability windows and therefore worse reliability. To reduce the exposure degree to failures, codewords should be spread across a shorter number of devices, which implies a shorter optimal codeword length. Also, lower values of the bandwidth constraint factor ϕ result in increased vulnerability windows, which in turn result in shorter optimal codeword lengths. In particular, we observe that the ratio of the optimal codeword lengths generally decreases with decreasing bandwidth constraint factor ϕ. However, when the optimal codeword lengths m^*_{exp} and m^*_{det} reach the value of the minimum codeword length, then the ratio becomes equal to 1, as shown in Figures 30(d) and 31(d) for the case of $k = 50$ and for $s_{\text{eff}} = 5/6, 6/7$ and $7/8$.

VII. Discussion

The symmetric and declustered data placement schemes reduce rebuild times by recovering data in parallel from the storage devices. In particular, for large-scale data storage systems, the rebuild times become extremely short. The model presented copes with this issue by considering the realistic case of network rebuild bandwidth constraints, which effectively prolong the duration of rebuild times.

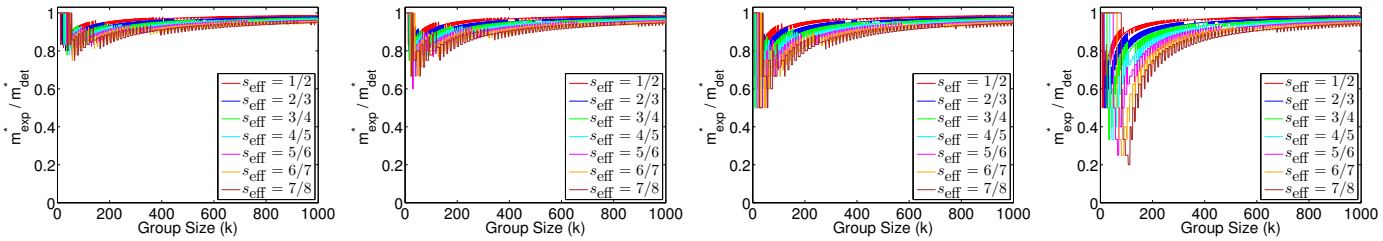
Although erasure coding schemes provide high data reliability and storage efficiency, the rebuild process involves I/O operations and network transfers that increase the consumption of device and network bandwidth. In particular, large MDS codes pose a challenge to the usage of network resources given that a lost symbol is recovered via an (m, l) erasure code by transferring a large number of l symbols from l surviving devices over the network. Although this may not be critical in purely archival tiers, recovering large amounts of data in active tiers results in additional traffic over increased time periods, which has an impact on the latency of the foreground workload and therefore affects system performance. This issue, also known as the *repair bandwidth problem*, has prompted the development of alternative erasure coding schemes that aim to reduce the amount of data transferred over the storage network during reconstruction (see [37][38] and references therein). They result in smaller amounts of data being read from the surviving devices and therefore in shorter rebuild times and higher reliabilities. However, in the case of *functional repairs*, a lost user-data symbol is replaced by an appropriate parity symbol, which now implies that reading such a user-data symbol can no longer be performed directly, but indirectly by accessing l symbols. Although this may be practical for archival tiers, it negatively affects the performance of the workloads encountered in active tiers. Therefore, emphasis is placed on *exact repairs* that preserve user data and maintain the erasure code in systematic form. The effect of these methods on system reliability is beyond the scope of this article and is a subject of further investigation.

The analytical findings of this work are relevant for the case of large erasure-coded data centers where a significant percentage of nodes fail each day [40]. Subsequently, the data recovery operations generate an excessive rebuild traffic that competes with the huge amount of traffic generated by the frequent access of a large number of storage devices [36]. To ensure a desired performance level, the network bandwidth devoted to the repair traffic must be contained. Furthermore,



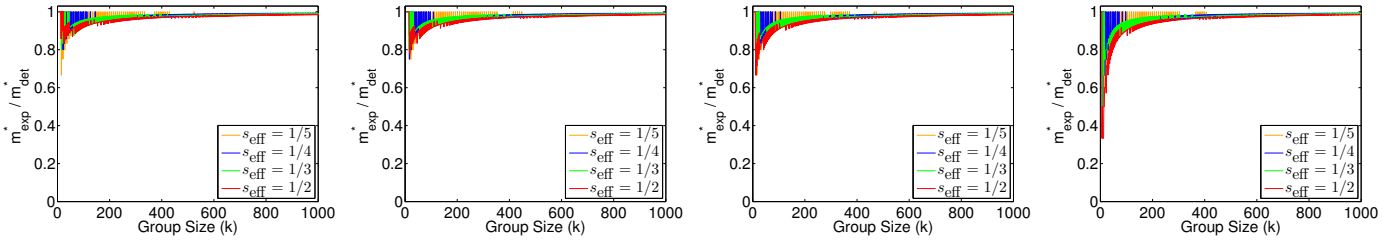
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 30. Ratio m^*_{exp} to m^*_{det} for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$.



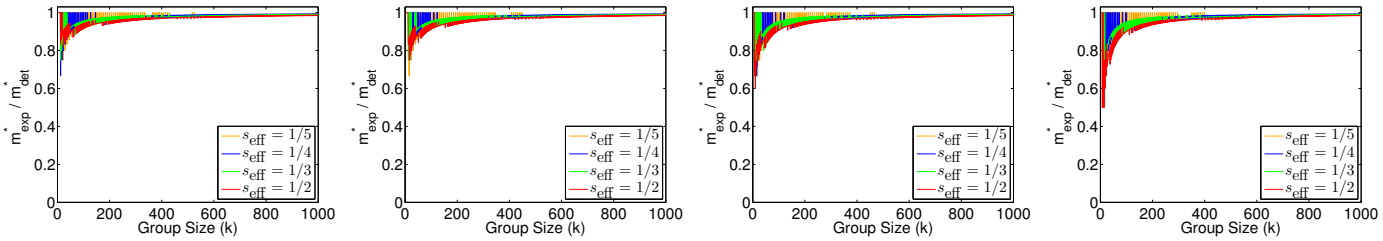
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 31. Ratio m^*_{exp} to m^*_{det} for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 32. Ratio m^*_{exp} to m^*_{det} for MTTDL vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 33. Ratio m^*_{exp} to m^*_{det} for EAFDL vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$.

for performance reasons, the codeword length should be kept relatively short, otherwise a large number of parity updates will interfere with the normal user traffic, resulting in a performance degradation. More specifically, Google's GFS as well as QFS use an RS(9,6) code that achieves a storage efficiency of 66% [32, 41], Facebook uses an RS(14,10) code that achieves a storage efficiency of 71% [34], and Windows Azure uses an LRC(16,10) code, which is not an MDS code, that achieves a storage efficiency of 75% [33]. Note that these systems initially used a three-way replication by storing three copies of all data, which achieved a storage efficiency of 33%. Consequently, to keep the storage overhead low, the erasure-code parameter values should be chosen such that the storage efficiency is in the range from 0.66 to 0.75.

VIII. Conclusions

Data storage systems use erasure coding schemes to recover lost data and enhance system reliability. However, network rebuild bandwidth constraints may degrade reliability. A general methodology was applied for deriving the Mean Time to Data Loss (MTTDL) and the Expected Annual Fraction of Data Loss (EAFDL) reliability metrics analytically. Closed-form expressions capturing the effect of a network rebuild

bandwidth constraint were obtained for the symmetric, clustered and declustered data placement schemes. We established that the reliability of storage systems is adversely affected by the network rebuild bandwidth constraints. The declustered placement scheme was found to offer superior reliability in terms of both metrics. We subsequently conducted an investigation of the reliability achieved by this scheme under various codeword configurations. The results demonstrated that both metrics are optimized by similar codeword lengths. For large storage systems that use a declustered placement scheme, the optimized codeword lengths are about 60% of the storage system size, independently of the network rebuild bandwidth constraints. The analytical reliability expressions derived can be used to identify redundancy and recovery schemes as well as data placement configurations that can achieve high reliability. The results can also be used to adapt the data placement schemes when the available network rebuild bandwidth or the number of devices in the system changes so that the system maintains a high level of reliability.

Extending the methodology developed to derive the reliability of erasure coded systems in the presence of unrecoverable latent errors is a subject of further investigation. Moreover, owing to the parallelism of the rebuild process, the model considered here yields very short rebuild times for large system sizes. Taking into account the fact that the rebuild times cannot be shorter than the actual failure detection times requires a more sophisticated modeling effort, which is also part of future work.

APPENDIX A
OPTIMAL $\hat{k_s}$ FOR MTTDL$^{\text{sym}}$

Proof of Proposition 1.

As mentioned in Section II-B, the system comprises n/k disjoint groups of k devices. We first obtain the optimal value for k by relaxing the constraint that n/k be an integer, that is, by considering all the integer values for k in the interval $I_k = [m+1, n]$. We subsequently impose the constraint and obtain the optimal value $\hat{k_s}$. Note that the constraint that n/k be an integer translates to $k \in I_k \cap D_n$, where D_n is the set of all integers that divide n, as defined in (44). Also, k_m, as defined in (45), represents the smallest integer in the interval $I_k = [m+1, n]$ that divides n. Thus, $k_m \in D_n$, $n \in I_k$, $n \in D_n$, and therefore $n \in I_k \cap D_n$.

Considering l, m, and n to be fixed, it follows from (39) that $\text{MTTDL}_k^{\text{sym}}$ is approximately proportional to the function A_k given by

$$A_k \triangleq \prod_{u=1}^{m-l} (k-u)^{m-l-u} \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right) . \quad (101)$$

Consequently, the value of k in the interval $[m+1, n]$ that maximizes $\text{MTTDL}_k^{\text{sym}}$ also maximizes A_k. Depending on the values of m and l, we consider the following two cases:

Case 1: $m - l = 1$. From (101) it follows that $A_k = \min\left(\frac{\phi}{1-\frac{1}{k}}, 1\right) \leq 1$, which is decreasing in k.

Depending on the value of ϕ, the following three subcases are considered:

(a) $\phi \geq 1 - \frac{1}{n} \Leftrightarrow n \leq \frac{1}{1-\phi}$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I_k = [m+1, n]$. Subsequently, imposing the constraint that n/k be an integer translates to $k \in I_a = I_k \cap D_n$. Note that I_a is not empty because $n \in I$ and $n \in D_n$. Furthermore, $k_m \in I_a$.

(b) $1 - \frac{1}{n} > \phi \geq 1 - \frac{1}{k_m} \Leftrightarrow k_m \leq \frac{1}{1-\phi} < n$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I = [m+1, \frac{1}{1-\phi}]$, which also includes k_m. Subsequently, imposing the constraint that n/k be an integer translates to $k \in I_b = I \cap D_n$. Note that I_b is not empty because $k_m \in I$ and $k_m \in D_n$, and therefore $k_m \in I_b$.

(c) $1 - \frac{1}{k_m} > \phi \geq 1 - \frac{1}{m+1} \Leftrightarrow m+1 \leq \frac{1}{1-\phi} < k_m$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I = [m+1, \frac{1}{1-\phi}]$, which does not include k_m. Consequently, none of the values in the interval I divide n. Subsequently, imposing the constraint that n/k be an integer translates to considering values of k that exceed k_m, in which case $A_k = \frac{\phi}{1-\frac{1}{k}} < 1$. As A_k is decreasing in k, we deduce that A_k is maximized when $k = k_m$.

(d) $\phi < 1 - \frac{1}{m+1} \Leftrightarrow m+1 > \frac{1}{1-\phi}$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{1}{k}} < \cdots < \frac{\phi}{1-\frac{1}{m+1}} < 1$, for $m+1 < k < n$. Therefore, $A_k = \frac{\phi}{1-\frac{1}{k}}$, $\forall\, k \in [m+1, n]$, and A_k is maximized when $k = m+1$. Subsequently, imposing the constraint that n/k be an integer translates to $k = k_m$.

Case 2: $m - l \geq 2$. It holds that $k > m = (m-l) + l \geq m - l + 1$ and, therefore, $k \geq m - l + 2$. From (101) it follows that

$$A_k = \left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u} \min(k\,\phi,\, k-u)\right] \min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) . \quad (102)$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi \geq 1 - \frac{m-l}{k} \Leftrightarrow k \leq \frac{m-l}{1-\phi}$. In this case, it holds that $\min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) = 1$, and

$$A_k = \prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u} \min(k\,\phi,\, k-u) . \quad (103)$$

Note that each of the terms in the product is increasing in k, which implies that A_k is also increasing in k, for $k \leq k_f$, where $k_f \triangleq \lfloor \frac{m-l}{1-\phi} \rfloor$.

(b) $\phi \leq 1 - \frac{m-l}{k} \Leftrightarrow k \geq \frac{m-l}{1-\phi}$. In this case, it holds that $\min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) = \frac{\phi}{1-\frac{m-l}{k}}$. Also, for $u < m-l$, it holds

that $k-u > k-(m-l) = k\,(1-\frac{m-l}{k}) \geq k\,\phi$, such that $\min(k\,\phi,\, k-u) = k\,\phi$. From (102), it now follows that

$$A_k = \left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u}\; k\,\phi\right] \frac{\phi}{1-\frac{m-l}{k}} = \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u}\right] \frac{{k}^{m-l}}{k-(m-l)}\,. \quad (104)$$

We proceed by recognizing that the last term in (104) is increasing in k, for $k \geq m-l+2$. This is a direct consequence of the fact that the function $f(x) = \frac{x^{m-l}}{x-(m-l)}$ is increasing in $x \in [\frac{(m-l)^2}{m-l-1}, \infty]$, and the fact that $m-l+2 \geq \frac{(m-l)^2}{m-l-1}$ for $m-l \geq 2$. Moreover, each of the terms in the product is increasing in k, which implies that A_k is also increasing in k, for $k \geq k_c$, where $k_c \triangleq \lceil \frac{m-l}{1-\phi} \rceil$.

To conclude that A_k is increasing in the entire range of k, it suffices to show that $A_{k_f} \leq A_{k_c}$. Clearly, this condition holds when $k_f = k_c$, that is, when $\frac{m-l}{1-\phi}$ is an integer. If $\frac{m-l}{1-\phi}$ is not an integer, then it holds that

$$k_f < \frac{m-l}{1-\phi} < k_c = k_f + 1\,, \quad (105)$$

which in turn implies that

$$1 - \frac{m-l}{k_f} < \phi < \phi_c \triangleq 1 - \frac{m-l}{k_c}\,. \quad (106)$$

Furthermore, using the relation $k_f \geq m-l+2 > m-l-1$, we deduce that $\frac{m-l-1}{k_f} < \frac{m-l}{k_f+1} = \frac{m-l}{k_c} = 1-\phi_c$, that is,

$$\phi_c < 1 - \frac{m-l-1}{k_f}\,. \quad (107)$$

For $u \leq m-l-1$, and combining (106) and (107) yields $k_f - u \geq k_f - (m-l-1) = k_f\,(1-\frac{m-l-1}{k_f}) > k_f\,\phi_c > k_f\,\phi$, such that $\min(k_f\,\phi,\, k_f - u) = k_f\,\phi$. Thus, (103) can be written as follows:

$$\begin{aligned} A_{k_f} &= \prod_{u=1}^{m-l-1} [(k_f-u)^{m-l-1-u} k_f\,\phi] \\ &= \phi^{m-l-1}\left[\prod_{u=1}^{m-l-1} (k_f-u)^{m-l-1-u}\right] {k_f}^{m-l-1} \\ &\overset{(106)}{<} \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k_f-u)^{m-l-1-u}\right] \frac{{k_f}^{m-l}}{k_f-(m-l)}\,. \end{aligned} \quad (108)$$

Also, from (104) we get

$$A_{k_c} = \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k_c-u)^{m-l-1-u}\right] \frac{{k_c}^{m-l}}{k_c-(m-l)}\,. \quad (109)$$

From (108) and (109), and given that $k_f < k_c$, it follows that $A_{k_f} < A_{k_c}$.

From the above, we conclude that when $m-l \geq 2$, A_k is increasing in k and, therefore, is maximized when $k = n$. ■

APPENDIX B
APPROXIMATE DERIVATION OF $\theta^{\,\mathrm{sym}}$

Proof of Lemma 1.

From (3) and (38), and using (49) and (50), it follows that

$$\log(\theta^{\,\mathrm{sym}}) = \sum_{u=1}^{hxk} \log\left(\min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right)\right)\,. \quad (110)$$

Given that $\frac{\phi}{1-\frac{u}{k}} \leq 1 \;\Leftrightarrow\; u \leq (1-\phi)k$, (110) yields

$$\begin{aligned} \log(\theta^{\,\mathrm{sym}}) &= \sum_{u=1}^{\widehat{\phi}k} \log\left(\frac{\phi}{1-\frac{u}{k}}\right) \\ &= \sum_{u=1}^{\widehat{\phi}k} \log(\phi) - \sum_{u=1}^{\widehat{\phi}k} \log\left(1-\frac{u}{k}\right) \\ &= \widehat{\phi}\,k\log(\phi) - \sum_{u=1}^{\widehat{\phi}k} \log\left(1-\frac{u}{k}\right)\,, \end{aligned} \quad (111)$$

where $\widehat{\phi} = \min(1-\phi, hx)$ as defined in (48). For large values of k, the preceding summation can be approximated using Lemma 1 of [25], which states that for small values of ϵ, that is, when ϵ approaches 0, and for any function $f(y)$, it holds that [25, Equation (122)]

$$\epsilon \sum_{j=1}^{\alpha/\epsilon} f(j\epsilon) \;\approx\; \int_{\frac{\epsilon}{2}}^{\alpha+\frac{\epsilon}{2}} f(y)\,dy\,, \qquad \forall\,\alpha \in \mathbb{R}\,. \quad (112)$$

For $\alpha = \widehat{\phi}$ and $f(y) = \log(1-y)$, (112) yields

$$\epsilon \sum_{j=1}^{\widehat{\phi}/\epsilon} \log(1-j\epsilon) \;\approx\; \int_{\frac{\epsilon}{2}}^{\widehat{\phi}+\frac{\epsilon}{2}} \log(1-y)\,dy\,. \quad (113)$$

Also, from Equations (128), (129), and (133) of [25], it follows that

$$\begin{aligned} \int_{\frac{\epsilon}{2}}^{\widehat{\phi}+\frac{\epsilon}{2}} \log(1-y)\,dy &= \log\left(\frac{(1-\frac{\epsilon}{2})^{1-\frac{\epsilon}{2}}}{(1-\widehat{\phi}-\frac{\epsilon}{2})^{1-\widehat{\phi}-\frac{\epsilon}{2}}}\right) - \widehat{\phi} \\ &\approx\; \log\left(\frac{1}{(1-\widehat{\phi})^{1-\widehat{\phi}}}\right) - \widehat{\phi} + \frac{\log(1-\widehat{\phi})}{2}\,\epsilon - \frac{\widehat{\phi}}{8(1-\widehat{\phi})}\,\epsilon^2\,. \end{aligned} \quad (114)$$

Substituting (114) into (113), and setting $\epsilon = 1/k$, yields

$$\begin{aligned} \sum_{j=1}^{\widehat{\phi}k} \log\left(1-\frac{j}{k}\right) \;\approx\; &-\left[\log\left((1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k \\ &+ \frac{1}{2}\log(1-\widehat{\phi}) - \frac{\widehat{\phi}}{8(1-\widehat{\phi})\,k}\,. \end{aligned} \quad (115)$$

Note that for large values of k, the last term of the right-hand side of (115) is negligible and therefore can be ignored. Substituting (115) into (111) yields the following approximation:

$$\begin{aligned} \log\left(\theta^{\,\mathrm{sym}}_{\,\mathrm{approx}}\right) \approx\; &\widehat{\phi}\,\log(\phi)\,k \\ &+ \left[\log\left((1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k - \frac{1}{2}\log(1-\widehat{\phi})\,, \end{aligned} \quad (116)$$

which yields (47). ■

Remark 14: When the average network rebuild bandwidth is upper limited by $B_{\max}$, the bandwidth constraint factor ϕ is no longer constant, but, for large values of k and according to (36), is given by

$$\phi = \frac{B_{\max}}{k\,b} \overset{(9)}{=} \frac{N_b}{k}\,, \tag{117}$$

which in turn implies that ϕ tends to 0 as k increases. Consequently, for large values of k, (48) yields

$$\widehat{\phi} = hx\,. \tag{118}$$

Substituting (117) and (118) into (116) yields

$$\begin{aligned}\log\left(\theta^{\text{sym}}_{\text{approx}}\right) \approx\ & hx\,\log\left(\frac{N_b}{k}\right)k \\ &+ \left[\log\left((1-hx)^{1-hx}\right) + hx\right]k - \frac{1}{2}\log(1-hx) \\ \approx\ & -hx\,k\,\log(k) + hx\log(N_b)\,k \\ &+ \left[\log\left((1-hx)^{1-hx}\right) + hx\right]k - \frac{1}{2}\log(1-hx)\,.\end{aligned} \tag{119}$$

Appendix C
Optimal $\hat{k}$ for MTTDL

Proof of Proposition 10.

Depending on the values of m and l, we consider the following three cases:

Case 1: $m-l=1$. Substituting $l=m-1$ into (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{m-1}{m}\cdot\frac{\min\left(\frac{\phi}{1-\frac{1}{\hat{k_s}}},1\right)}{\min\left(\frac{\phi}{1-\frac{1}{m}},1\right)} < 1\,. \tag{120}$$

The inequality holds because $\frac{m-1}{m}<1$ and $\min\left(\frac{\phi}{1-\frac{1}{\hat{k_s}}},1\right) \leq \min\left(\frac{\phi}{1-\frac{1}{m}},1\right)$, given that $\hat{k_s} \geq m+1 > m$ and, therefore, $\frac{\phi}{1-\frac{1}{\hat{k_s}}} < \frac{\phi}{1-\frac{1}{m}}$. Consequently, the MTTDL is maximized by the clustered placement scheme.

Case 2: $m-l=2$. This implies that $n/2 \geq m = l+2 \geq 3$. Consequently, $1 \leq \frac{2(m-2)}{(m-1)} = \frac{(m-2)(2m-2)}{(m-1)^2} \leq \frac{(m-2)(n-2)}{(m-1)^2} = \frac{(m-2)n^2}{(n-2)(m-1)^2}\left(1-\frac{2}{n}\right)^2$, which in turn implies that

$$G \leq 1-\frac{2}{n} < 1-\frac{1}{n}\,, \tag{121}$$

where

$$G \triangleq G(m,n) = \sqrt{\frac{n-2}{m-2}}\ \frac{m-1}{n}\,. \tag{122}$$

For $m-l=2$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k_s}} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-2)(n-1)\min\left(\frac{\phi}{1-\frac{1}{n}},1\right)\min\left(\frac{\phi}{1-\frac{2}{n}},1\right)}{(m-1)^2\ \min\left(\frac{\phi}{1-\frac{2}{m}},1\right)^2}\,. \tag{123}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > G$. In this case, it holds that $\frac{G}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{n}}$ and $\frac{G}{1-\frac{2}{n}} < \frac{\phi}{1-\frac{2}{n}}$. Also, it holds from (121) that $\frac{G}{1-\frac{1}{n}} < \frac{G}{1-\frac{2}{n}} \leq 1$. Consequently, $\min\left(\frac{\phi}{1-\frac{1}{n}},1\right) > \frac{G}{1-\frac{1}{n}}$, and $\min\left(\frac{\phi}{1-\frac{2}{n}},1\right) \geq \frac{G}{1-\frac{2}{n}}$. Moreover, from (123), and using the fact that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right) \leq 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \frac{(m-2)(n-1)\ \frac{G}{1-\frac{1}{n}}\ \frac{G}{1-\frac{2}{n}}}{(m-1)^2} \overset{(122)}{=} 1\,. \tag{124}$$

(b) $\phi \leq G$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{2}{n}} \leq \frac{\phi}{G} \leq 1$. It follows from (123) that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-2)(n-1)\ \frac{\phi}{1-\frac{1}{n}}\ \frac{\phi}{1-\frac{2}{n}}}{(m-1)^2\ \min\left(\frac{\phi}{1-\frac{2}{m}},1\right)^2}\,. \tag{125}$$

Depending on the value of ϕ, the following two subcases are considered:

(i) $\phi \geq 1-\frac{2}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right)=1$, and from (125) it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < \frac{(m-2)(n-1)\ \frac{G}{1-\frac{1}{n}}\ \frac{G}{1-\frac{2}{n}}}{(m-1)^2} \overset{(122)}{=} 1\,. \tag{126}$$

Also, in this case it holds that $1-\frac{2}{m} \leq G$. Note that for $n/2 \geq m$, it holds that

$$\left(\frac{1-\frac{2}{m}}{G}\right)^2 = \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)} \geq \frac{2\,(m-2)^3}{(m-1)^3}\,. \tag{127}$$

We now deduce that $m \leq 5$, because for $m \geq 6$, it holds that $\frac{2\,(m-2)^3}{(m-1)^3} > 1$ and, therefore, $1-\frac{2}{m} > G$, which is a contradiction. It turns out that for $m=3$ and $m=4$, the ratio $(1-\frac{2}{m})/G$ is less than 1 or, equivalently, $1-\frac{2}{m} < G$, when $n \leq 33$ and $n \leq 12$, respectively. For $m=5$, this ratio is less than 1 when $n=10$.

(ii) $\phi < 1-\frac{2}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right) = \frac{\phi}{1-\frac{2}{m}}$, and (125) yields

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\approx \frac{(m-2)(n-1)\ \frac{\phi}{1-\frac{1}{n}}\ \frac{\phi}{1-\frac{2}{n}}}{(m-1)^2\ \left(\frac{\phi}{1-\frac{2}{m}}\right)^2} \\ &= \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)} \overset{(122)}{=} \left(\frac{1-\frac{2}{m}}{G}\right)^2\,. \end{aligned} \tag{128}$$

As argued above, it holds for $m \geq 6$ that $1-\frac{2}{m} > G$ and, therefore, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > 1$. For $3 \leq m \leq 5$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ may be less than 1. In particular, for $m=3$ and $m=4$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ is less than 1 when $n \leq 33$ and $n \leq 12$, respectively. For $m=5$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ is less than 1 when $n=10$.

From the results obtained in the preceding two subcases, we conclude that, when $m-l=2$, the MTTDL is maximized

by the clustered placement scheme ($r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$) only in the following cases:

1) $m = 3$, $n = 3\,j$ with $2 \le j \le 11$, and $\phi < G = \frac{2\sqrt{n-2}}{n}$,
2) $m = 4$, $n = 8$, and $\phi < G = 3\sqrt{3}/8 = 0.649$,
3) $m = 4$, $n = 12$, and $\phi < G = \sqrt{5}/4 = 0.559$, and
4) $m = 5$, $n = 10$, and $\phi < G = 4\sqrt{2}/(5\sqrt{3}) = 0.653$.

In these cases, $\hat{k} = m$, whereas in all other cases, the MTTDL is maximized by the declustered placement scheme ($\hat{k} = n$).

Case 3: $m-l = 3$. This implies that $n/2 \ge m = l+3 \ge 4$. Thus, $n - 3 > m - 2$ or $(n-3)^2/(m-2)^2 > 1$. Also, for $m \ge 4$, it holds that $\frac{(m-3)(2m-1)}{(m-1)(m-2)} > 1$. Consequently, $1 < \frac{(n-3)^2(m-3)(2m-1)}{(m-2)^2(m-1)(m-2)} \le \frac{(n-3)^2(m-3)(n-1)}{(m-1)(m-2)^3} = \frac{(m-3)(n-1)n^3}{(n-3)(m-1)(m-2)^3}\left(1-\frac{3}{n}\right)^3$, which in turn implies that

$$Q < 1 - \frac{3}{n} < 1 - \frac{2}{n} < 1 - \frac{1}{n}\,, \tag{129}$$

where

$$Q \triangleq Q(m,n) = \sqrt[3]{\frac{(n-3)(m-1)}{(m-3)(n-1)}}\;\frac{m-2}{n}\,. \tag{130}$$

For $m-l = 3$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k_s}} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)}{(m-1)(m-2)^3} \cdot \frac{\min\left(\frac{\phi}{1-\frac{1}{n}},1\right)\min\left(\frac{\phi}{1-\frac{2}{n}},1\right)\min\left(\frac{\phi}{1-\frac{3}{n}},1\right)}{\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3}\,. \tag{131}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > Q$. In this case, it holds that $\frac{Q}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1,2,3$. Also, it holds from (129) that $\frac{Q}{1-\frac{u}{n}} < 1$, for $u = 1,2,3$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}},1\right)$, for $u = 1,2,3$. Moreover, from (131), and using the fact that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) < 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \frac{(m-3)(n-1)^2(n-2)\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{3}{n}}}{(m-1)(m-2)^3} \overset{(130)}{=} 1\,. \tag{132}$$

(b) $\phi \le Q$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{2}{n}} < \frac{\phi}{1-\frac{3}{n}} < \frac{\phi}{Q} < 1$. It follows from (131) that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)\;\frac{\phi}{1-\frac{1}{n}}\;\frac{\phi}{1-\frac{2}{n}}\;\frac{\phi}{1-\frac{3}{n}}}{(m-1)(m-2)^3\;\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3}\,. \tag{133}$$

Depending on the value of ϕ, the following two subcases are considered:

(i) $\phi \ge 1-\frac{3}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) = 1$, and from (133) it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < \frac{(m-3)(n-1)^2(n-2)\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{2}{n}}\;\frac{Q}{1-\frac{3}{n}}}{(m-1)(m-2)^3} \overset{(130)}{=} 1\,. \tag{134}$$

Also, in this case it holds that $1-\frac{3}{m} \le Q$. Note that for $n/2 \ge m$, it holds that

$$\left(\frac{1-\frac{3}{m}}{Q}\right)^3 = \frac{(m-3)^4\,(n-1)\,n^3}{m^3\,(m-1)\,(m-2)^3\,(n-3)} \ge \left[4\left(\frac{m-3}{m-2}\right)^3\right]\left[\frac{2(m-3)(2m-1)}{(m-1)(2m-3)}\right]\,. \tag{135}$$

We now deduce that $m = 4$, because for $m \ge 5$, each of the terms in the two brackets is greater than one and, therefore, $1-\frac{3}{m} > Q$, which is a contradiction. It turns out that for $m = 4$, the ratio $(1-\frac{3}{m})/Q$ is less than one or, equivalently, $1-\frac{3}{m} < Q$, only when $n = 8$.

(ii) $\phi < 1-\frac{3}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) = \frac{\phi}{1-\frac{3}{m}}$, and (133) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)\;\frac{\phi}{1-\frac{1}{n}}\;\frac{\phi}{1-\frac{2}{n}}\;\frac{\phi}{1-\frac{3}{n}}}{(m-1)(m-2)^3\;\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3} = \left(\frac{1-\frac{3}{m}}{Q}\right)^3 \overset{(130)}{=} \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)}\,. \tag{136}$$

As argued above, it holds for $m \ge 5$ that $1-\frac{3}{m} > Q$ and, therefore, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > 1$. In fact, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$ only when $m = 4$ and $n = 8$.

From the results obtained in the preceding two subcases, we conclude that, when $m-l = 3$, MTTDL is maximized by the clustered placement scheme ($r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$) only when $l = 1$, $m = 4$, $n = 8$, and $\phi < Q = \sqrt[3]{15}/(4\sqrt[3]{7}) = 0.322$. In all other cases, MTTDL is maximized by the declustered placement scheme.

Case 4: $m-l \ge 4$. We deduce from (89) that $n-m \ge m$ and

$$n-m+l \ge m+l = m-l+2l \ge 4+2l \ge 6\,. \tag{137}$$

Let us define

$$R_M \triangleq R_M(l,m,n) = \frac{l+1}{n}\left[\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u-1}\right]^{\frac{1}{m-l}}\,. \tag{138}$$

Next, we will show that

$$S_M \triangleq \left(\frac{1-\frac{m-l}{m}}{R_M}\right)^{m-l} = \left(\frac{l}{m\,R_M}\right)^{m-l} > 1\,. \tag{139}$$

Substituting (138) into (139) yields

$$S_M = \left[\frac{l\,n}{(l+1)\,m}\right]^{m-l}\;\prod_{u=1}^{m-l}\left(\frac{n-u}{m-u}\right)^{m-l-u-1}\,. \tag{140}$$

It follows from (89) that $\frac{l}{l+1} \geq \frac{1}{2}$ and $\frac{n}{m} \geq 2$. Consequently, the term in brackets is greater than or equal to 1. Next, we will show that the product is greater than 1. For $m-l \geq 4$, the product can be written as follows:

$$\prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} = \prod_{u=1}^{m-l-4} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \prod_{u=m-l-3}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} . \tag{141}$$

Clearly, for $n > m$, the first product is greater than or equal to 1. The second product is greater than 1 because it can be written as follows:

$$\begin{aligned} &\prod_{u=m-l-3}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \\ &= \left[\frac{n-(m-l-3)}{m-(m-l-3)}\right]^2 \cdot \frac{n-(m-l-2)}{m-(m-l-2)} \cdot \frac{m-(m-l)}{n-(m-l)} \\ &= 1 + \frac{(n-m)[l(n-m+l)(n-m+2l+8)-18]}{(l+3)(l+2)(n-m+l)} \\ &\overset{(89)(137)}{\geq} 1 + \frac{(n-m)[6\,(6+l+8)-18]}{(l+3)(l+2)(n-m+l)} > 1 . \end{aligned} \tag{142}$$

Inequality (139) is a direct consequence of (140), (141), and (142).

We deduce from (139) that

$$R_M < \frac{l}{m} = 1 - \frac{m-l}{m} < 1 - \frac{m-l}{n} < \cdots < 1 - \frac{1}{n} . \tag{143}$$

For $m-l \geq 4$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k}_s} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) . \tag{144}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > R_M$. In this case, it holds that $\frac{R_M}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Also, it holds from (143) that $\frac{R_M}{1-\frac{u}{n}} < 1$, for $u = 1, \ldots, m-l$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) > \frac{R_M}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Moreover, from (144), and using the fact that $\min\left(\frac{\phi}{1-\frac{m-l}{m}}, 1\right) \leq 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \frac{R_M}{1-\frac{u}{n}} \overset{(138)}{=} 1 . \tag{145}$$

(b) $\phi \leq R_M$. In this case, it follows from (143) that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{m-l}{n}} < \frac{\phi}{1-\frac{m-l}{m}} = \frac{m\,\phi}{l} < \frac{\phi}{R_M} \leq 1$. Subsequently, (144) yields

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \left(\frac{m\,\phi}{l}\right)^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \left(\frac{\phi}{1-\frac{u}{n}}\right) \\ &= \left[\frac{l\,n}{(l+1)\,m}\right]^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \\ &\overset{(138)}{=} \left(\frac{l}{m\,R_M}\right)^{m-l} \overset{(139)}{>} 1 . \end{aligned} \tag{146}$$

From the results obtained in the preceding two subcases, we conclude that, when $m-l \geq 4$, MTTDL is maximized by the declustered placement scheme. ■

APPENDIX D
OPTIMAL k FOR EAFDL

Proof of Proposition 11.

Depending on the values of m and l, we consider the following two cases:

Case 1: $m-l = 1$. In this case, (92) yields

$$r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} \approx \frac{m}{n-1} \cdot \frac{\min\left(\frac{\phi}{1-\frac{1}{m}}, 1\right)}{\min\left(\frac{\phi}{1-\frac{1}{n}}, 1\right)} . \tag{147}$$

It follows from (89) that $\frac{l+1}{l} \leq 2$ and $\frac{m}{n} \leq \frac{1}{2}$, with the equalities holding only when $l = 1$, $m = 2$, and $n = 4$. Consequently,

$$\frac{(l+1)\,m}{l\,n} = \frac{m^2}{(m-1)\,n} \leq 1 , \tag{148}$$

with the equality holding only when $l+1 = m = 2$ and $n = 4$.

We deduce from (148) that

$$\frac{m}{n} \leq \frac{m-1}{m} = 1 - \frac{1}{m} < 1 - \frac{1}{n} . \tag{149}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > \frac{m}{n}$. In this case, it holds that $\frac{\frac{m}{n}}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{n}}$. Also, it holds from (149) that $\frac{\frac{m}{n}}{1-\frac{1}{n}} < 1$. Consequently, $\min\left(\frac{\phi}{1-\frac{1}{n}}, 1\right) > \frac{\frac{m}{n}}{1-\frac{1}{n}}$. Moreover, from (147), and using the fact that $\min\left(\frac{\phi}{1-\frac{1}{m}}, 1\right) \leq 1$, it follows that

$$r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} < \frac{m}{n-1} \cdot \frac{1}{\frac{\frac{m}{n}}{1-\frac{1}{n}}} = 1 . \tag{150}$$

(b) $\phi \leq \frac{m}{n}$. In this case, it follows from (149) that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{m}} \leq \frac{\phi}{\frac{m}{n}} \leq 1$. Subsequently, (147) yields

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} \approx \frac{m}{n-1} \cdot \frac{\frac{\phi}{1-\frac{1}{m}}}{\frac{\phi}{1-\frac{1}{n}}} = \frac{m^2}{(m-1)\,n} \;\leq\; 1\,, \quad (151)$$

with the equality holding only when $l = 1$, $m = 2$, and $n = 4$. In this case, the clustered and declustered placements yield the same reliability.

From the results obtained in the preceding two subcases, we conclude that, when $m - l = 1$, EAFDL is minimized by the declustered placement scheme.

Case 2: $m - l \geq 2$. Let us define

$$R_E \triangleq R_E(l,m,n) = \frac{l+1}{n}\left[\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u}\right]^{\frac{1}{m-l}}. \quad (152)$$

Note that the following inequality holds

$$\left(\frac{R_E}{1-\frac{m-l}{m}}\right)^{m-l} = \left(\frac{m\,R_E}{l}\right)^{m-l} \overset{(152)}{=} \left[\frac{(l+1)\,m}{l\,n}\right]^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u} < 1\,, \quad (153)$$

because it holds that $(l+1)\,m/(l\,n) \leq 1$, as shown in Appendix C, and $(m-u)/(n-u) < 1$, for $u = 1, \ldots, m-l-1$, owing to (89).

We deduce from (153) that

$$R_E < \frac{l}{m} = 1 - \frac{m-l}{m} < 1 - \frac{m-l}{n} < \cdots < 1 - \frac{1}{n}\,. \quad (154)$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > R_E$. In this case, it holds that $\frac{R_E}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Also, it holds from (154) that $\frac{R_E}{1-\frac{u}{n}} < 1$ for $u = 1, \ldots, m-l$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) > \frac{R_E}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Moreover, from (92), and using the fact that $\min\left(\frac{\phi}{1-\frac{m-l}{m}}, 1\right) \leq 1$, it follows that

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} < (l+1)^{m-l}\,\frac{(l-1)!}{(m-1)!}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Big/ \frac{R_E}{1-\frac{u}{n}} \overset{(152)}{=} 1\,. \quad (155)$$

(b) $\phi \leq R_E$. It follows from (154) that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{m-l}{n}} < \frac{\phi}{1-\frac{m-l}{m}} = \frac{m\,\phi}{l} \leq \frac{\phi}{R_E} \leq 1$. Subsequently, (92) yields

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} \approx (l+1)^{m-l}\,\frac{(l-1)!}{(m-1)!}\left(\frac{m\,\phi}{l}\right)^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Big/ \left(\frac{\phi}{1-\frac{u}{n}}\right) = \left[\frac{(l+1)\,m}{l\,n}\right]^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u} \overset{(152)}{=} \left(\frac{m\,R_E}{l}\right)^{m-l} \overset{(153)}{<} 1\,. \quad (156)$$

From the results obtained in the preceding two subcases, we conclude that, when $m - l \geq 2$, EAFDL is minimized by the declustered placement scheme. ■

REFERENCES

[1] I. Iliadis, "Reliability of erasure coded systems under rebuild bandwidth constraints," in Proceedings of the 11th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2018, pp. 1–10.

[2] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 1988, pp. 109–116.

[3] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson, "RAID: High-performance, reliable secondary storage," ACM Comput. Surv., vol. 26, no. 2, Jun. 1994, pp. 145–185.

[4] M. Malhotra and K. S. Trivedi, "Reliability analysis of redundant arrays of inexpensive disks," J. Parallel Distrib. Comput., vol. 17, Jan. 1993, pp. 146–151.

[5] W. A. Burkhard and J. Menon, "Disk array storage system reliability," in Proceedings of the 23rd International Symposium on Fault-Tolerant Computing, Jun. 1993, pp. 432–441.

[6] K. S. Trivedi, Probabilistic and Statistics with Reliability, Queueing and Computer Science Applications, 2nd ed. New York: Wiley, 2002.

[7] Q. Xin, E. L. Miller, T. J. E. Schwarz, D. D. E. Long, S. A. Brandt, and W. Litwin, "Reliability mechanisms for very large storage systems," in Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST), Apr. 2003, pp. 146–156.

[8] T. J. E. Schwarz, Q. Xin, E. L. Miller, D. D. E. Long, A. Hospodor, and S. Ng, "Disk scrubbing in large archival storage systems," in Proceedings of the 12th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Oct. 2004, pp. 409–418.

[9] Q. Lian, W. Chen, and Z. Zhang, "On the impact of replica placement to the reliability of distributed brick storage systems," in Proc. 25th IEEE International Conference on Distributed Computing Systems (ICDCS), Jun. 2005, pp. 187–196.

[10] S. Ramabhadran and J. Pasquale, "Analysis of long-running replicated systems," in Proc. 25th IEEE International Conference on Computer Communications (INFOCOM), Apr. 2006, pp. 1–9.

[11] B. Eckart, X. Chen, X. He, and S. L. Scott, "Failure prediction models for proactive fault tolerance within storage systems," in Proceedings of the 16th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2008, pp. 1–8.

[12] A. Thomasian and M. Blaum, "Higher reliability redundant disk arrays: Organization, operation, and coding," ACM Trans. Storage, vol. 5, no. 3, Nov. 2009, pp. 1–59.

[13] K. Rao, J. L. Hafner, and R. A. Golding, "Reliability for networked storage nodes," IEEE Trans. Dependable Secure Comput., vol. 8, no. 3, May 2011, pp. 404–418.

[14] I. Iliadis, R. Haas, X.-Y. Hu, and E. Eleftheriou, "Disk scrubbing versus intradisk redundancy for RAID storage systems," ACM Trans. Storage, vol. 7, no. 2, Jul. 2011, pp. 1–42.

[15] V. Venkatesan, I. Iliadis, C. Fragouli, and R. Urbanke, "Reliability of clustered vs. declustered replica placement in data storage systems," in Proceedings of the 19th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Jul. 2011, pp. 307–317.

[16] V. Venkatesan, I. Iliadis, and R. Haas, "Reliability of data storage systems under network rebuild bandwidth constraints," in Proceedings of the 20th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Aug. 2012, pp. 189–197.

[17] V. Venkatesan and I. Iliadis, "A general reliability model for data storage systems," in Proceedings of the 9th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2012, pp. 209–219.

[18] J.-F. Pâris, T. J. E. Schwarz, A. Amer, and D. D. E. Long, "Highly reliable two-dimensional RAID arrays for archival storage," in Proceedings of the 31st IEEE International Performance Computing and Communications Conference (IPCCC), Dec. 2012, pp. 324–331.

[19] V. Venkatesan and I. Iliadis, "Effect of codeword placement on the reliability of erasure coded data storage systems," in Proceedings of the 10th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2013, pp. 241–257.

[20] I. Iliadis and V. Venkatesan, "An efficient method for reliability evaluation of data storage systems," in Proceedings of the 8th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2015, pp. 6–12.

[21] ——, "Most probable paths to data loss: An efficient method for reliability evaluation of data storage systems," Int'l J. Adv. Syst. Measur., vol. 8, no. 3&4, Dec. 2015, pp. 178–200.

[22] S. Caron, F. Giroire, D. Mazauric, J. Monteiro, and S. Pérennes, "P2P storage systems: Study of different placement policies," Peer-to-Peer Networking and Applications, Mar. 2013, pp. 1–17.

[23] I. Iliadis and V. Venkatesan, "Expected annual fraction of data loss as a metric for data storage reliability," in Proceedings of the 22nd Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2014, pp. 375–384.

[24] ——, "Reliability assessment of erasure coded systems," in Proceedings of the 10th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2017, pp. 41–50.

[25] ——, "Reliability evaluation of erasure coded systems," Int'l J. Adv. Telecommun., vol. 10, no. 3&4, Dec. 2017, pp. 118–144.

[26] J. G. Elerath and J. Schindler, "Beyond MTTDL: A closed-form RAID 6 reliability equation," ACM Trans. Storage, vol. 10, no. 2, Mar. 2014, pp. 1–21.

[27] I. Iliadis and V. Venkatesan, "Rebuttal to 'Beyond MTTDL: A closed-form RAID-6 reliability equation'," ACM Trans. Storage, vol. 11, no. 2, Mar. 2015, pp. 1–10.

[28] "Amazon Simple Storage Service." [Online]. Available: http://aws.amazon.com/s3/ [retrieved: November 2017]

[29] D. Borthakur et al., "Apache Hadoop goes realtime at Facebook," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 2011, pp. 1071–1080.

[30] R. J. Chansler, "Data availability and durability with the Hadoop Distributed File System," ;login: The USENIX Association Newsletter, vol. 37, no. 1, 2013, pp. 16–22.

[31] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies (MSST), May 2010, pp. 1–10.

[32] D. Ford et al., "Availability in globally distributed storage systems," in Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2010, pp. 61–74.

[33] C. Huang et al., "Erasure coding in Windows Azure Storage," in Proceedings of the USENIX Annual Technical Conference (ATC), Jun. 2012, pp. 15–26.

[34] S. Muralidhar et al., "f4: Facebook's Warm BLOB Storage System," in Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2014, pp. 383–397.

[35] "IBM Cloud Object Storage." [Online]. Available: www.ibm.com/cloud-computing/products/storage/object-storage/how-it-works/ [retrieved: November 2017]

[36] K. V. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K. Ramchandran, "A solution to the network challenges of data recovery in erasure-coded distributed storage systems: A study on the Facebook warehouse cluster," in Proceedings of the 5th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage), Jun. 2013, pp. 1–5.

[37] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network coding for distributed storage," Proc. IEEE, vol. 99, no. 3, Mar. 2011, pp. 476–489.

[38] M. Zhang, S. Han, and P. P. C. Lee, "A simulation analysis of reliability in erasure-coded data centers," in Proceedings of the 36th IEEE Symposium on Reliable Distributed Systems (SRDS), Sep. 2017, pp. 144–153.

[39] J. E. Angus, "On computing MTBF for a k-out-of-n:G repairable system," IEEE Trans. Reliability, vol. 37, no. 3, Aug. 1988, pp. 312–313.

[40] M. Silberstein, L. Ganesh, Y. Wang, L. Alvisi, and M. Dahlin, "Lazy means smart: Reducing repair bandwidth costs in erasure-coded distributed storage," in Proceedings of the 7th ACM International Systems and Storage Conference (SYSTOR), Jun. 2014, pp. 15:1–15:7.

[41] M. Ovsiannikov et al., "The quantcast file system," in Proc. 39th Int'l Conf. on Very Large Data Bases (VLDB), vol. 6, no. 11. VLDB Endowment, Aug. 2013, pp. 1092–1101.