Most Probable Paths to Data Loss: An Efficient Method for Reliability Evaluation of Data Storage Systems

Ilias Iliadis and Vinodh Venkatesan
IBM Research – Zurich
Email: {ili,ven}@zurich.ibm.com

Abstract—**The effectiveness of the redundancy schemes that have been developed to enhance the reliability of storage systems has predominantly been evaluated based on the mean time to data loss (**MTTDL**) metric. This metric has been widely used to compare schemes, to assess tradeoffs, and to estimate the effect of various parameters on system reliability. Analytical expressions for** MTTDL **are typically derived using Markov chain models. Such derivations, however, remain a challenging task owing to the high complexity of the analysis of the Markov chains involved, and therefore the system reliability is often assessed by rough approximations. To address this issue, a general methodology based on the direct-path approximation was used to obtain the** MTTDL **analytically for a class of redundancy schemes and for failure time distributions that also include real-world distributions, such as Weibull and gamma. The methodology, however, was developed for the case of a single direct path to data loss. This work establishes that this methodology can be extended and used in the case where there are multiple shortest paths to data loss to approximately derive the** MTTDL **for a broader set of redundancy schemes. The value of this simple, yet efficient methodology is demonstrated in several contexts. It is verified that the results obtained for RAID-5 and RAID-6 systems match with those obtained in previous work. As a further demonstration, we derive the exact** MTTDL **of a specific RAID-51 system and confirm that it matches with the** MTTDL **obtained from the methodology proposed. In some cases, the shortest paths are not necessarily the most probable ones. We establish that this methodology can be extended to the most probable paths to data loss to derive closed-form approximations for the** MTTDL **of RAID-6 and two-dimensional RAID-5 systems in the presence of unrecoverable errors and device failures. A thorough comparison of the reliability level achieved by the redundancy schemes considered is also conducted.**

Keywords–Shortest path; direct path; data loss; latent errors; MTTDL; rebuild; rare events; RAID; closed-form; analysis.

I. Introduction

Storage systems experience data losses due to device failures, including disk and node failures. To avoid a permanent loss of data, redundancy schemes were developed that enable the recovery of this data. However, during rebuild operations, additional device failures may occur that eventually lead to permanent data losses. There is a variety of redundancy schemes that offer different levels of reliability as they tolerate varying degrees of device failures. Each of these schemes is characterized by an overhead, which reflects the additional operations that need to be performed for maintaining data consistency, and a storage efficiency, which expresses the additional amount of data, referred to as parity, that needs to be stored in the system.

The reliability of storage systems and the effectiveness of redundancy schemes have predominantly been assessed based on the mean time to data loss (MTTDL) metric, which expresses the amount of time that is expected to elapse until the first data is irrecoverably lost [1][2][3]. During this period, failures cause data to be temporarily lost, which is subsequently recovered owing to the redundancy built into the system.

Analytical expressions for the MTTDL are typically derived using Markov chain models [4], which assume that the times to component failures are independent and exponentially distributed. A methodology for obtaining MTTDL under general non-exponential failure and rebuild time distributions, which therefore does not involve any Markov analysis, was presented in [5]. The complexity of these derivations depends on the redundancy schemes and the underlying system configurations considered. The MTTDL metric has been proven useful for assessing tradeoffs, for comparing schemes, and for estimating the effect of various parameters on system reliability [6][7][8][9]. Analytical closed-form expressions for the MTTDL provide an accurate account of the effect of various parameters on system reliability. However, deriving exact closed-form expressions remains a challenging task owing to the high complexity of the analysis of the Markov chains involved [10][11]. For this reason, the system reliability is often assessed by rough approximations. As the direct MTTDL analysis is typically hard, an alternative is performing event-driven simulations [12][13]. However, simulations do not provide insight into how the various parameters affect the system reliability. This article addresses these issues by presenting a simple, yet efficient method, referred to as *most-probable-path approximation*, to obtain the MTTDL analytically for a broad set of redundancy schemes. It achieves that by considering the most likely paths that lead to data loss, which are the shortest ones. In contrast to simulations, this method provides approximate closed-form expressions for the MTTDL, thus circumventing the inherent complexity of deriving exact expressions using Markov analysis. Note also that this method was previously applied in the context of assessing system unavailability, in particular for systems characterized by large Markov chains [14]. It turns out that this approach agrees with the principle encountered in the probability context expressed by the phrase *“rare events occur in the most likely way”*. This is also demonstrated in [15], where the reliability level of systems composed of highly reliable components is essentially determined by the so-called "main event", which is the shortest way of failure appearance, that is, along the minimal monotone paths.

In [5][16][17][18][19], it was shown that the direct-path approximation, which considers paths without loops, yields accurate analytical reliability results. To further investigate the validity of the shortest-path-approximation method, we apply it to derive the MTTDL results for RAID-5 and RAID-6 systems and subsequently verify that they match with those obtained in previous works [2][3] for practical cases where the device failure rates are much smaller than the device rebuild rates. In all these previous works though, there is a single direct path to data loss. In contrast, our article is concerned with the case where there are multiple shortest paths to data loss. In this work, we investigate this issue and establish that the shortest-path-approximation method can be extended and also applied in the case of multiple shortest paths and can yield accurate reliability results. In particular, we derive the approximate MTTDL of a RAID-51 system using the shortest-path approximation. Subsequently, as a demonstration of the validity of the method proposed, we derive the exact MTTDL for a specific instance of a RAID-51 system and confirm that it matches with the corresponding MTTDL obtained using our method. Furthermore, we establish that the shortest-path approximation can be extended to the most probable path approximation in cases where the shortest paths may not necessarily be the most probable ones. In fact, an approximation that considers all direct paths implicitly considers the most probable ones because the direct paths are the most probable ones owing to the absence of loops.

The key contributions of this article are the following. We consider the reliability of the RAID-5, RAID-6, and RAID-51 systems that was assessed in our earlier work [1]. In this study, we extend our previous work by also considering two-dimensional RAID-5 systems. The MTTDL of a specific square two-dimensional RAID-5 system was estimated through a Markov chain model in [20], but no closed-form expression was provided owing to its complexity. In this work, using the shortest-path-approximation method, we obtain approximate closed-form expressions for the MTTDL that are general, simple, yet accurate for real-world systems. Furthermore, we perform a thorough comparison of the reliability levels, in terms of the MTTDL, achieved by these schemes. Subsequently, we consider the reliability of RAID-6 and two-dimensional RAID-5 systems in the presence of unrecoverable (latent) errors and device failures, and establish that in general the shortest paths may not be the most probable ones, A new enhanced methodology that considers the most probable paths, as opposed to the shortest paths, is subsequently introduced for efficiently assessing system reliability.

The remainder of the paper is organized as follows. Section II reviews the general framework for deriving the MTTDL of a storage system. Subsequently, the notion of the direct path to data loss is discussed in Section III, and the efficiency of the direct-path approximation is demonstrated in Section IV. Section V discusses the case of multiple shortest paths to data loss and presents the analysis of the RAID-51 and two-dimensional RAID-5 systems. Section VI presents a thorough comparison of the various redundancy schemes considered. Section VII provides a detailed analysis and comparison of the RAID-6 and two-dimensional RAID-5 systems in the presence of independent unrecoverable sector errors. The shortest-path-approximation method is enhanced to account for the most probable paths. Finally, we conclude in Section IX.

II. Derivation of MTTDL

In this section, we review the various methods that are used to obtain the MTTDL analytically.

A. Markov Analysis

Continuous-time Markov chain (CTMC) models reflecting the system operation can be constructed when the device failures and rebuild times are assumed to be independent and exponentially distributed. Under these assumptions, an appropriate CTMC model can be formulated to characterize the system behavior and capture the corresponding state transitions, including those that lead to data loss. Subsequently, using the infinitesimal generator matrix approach and determining the average time spent in the transient states of the Markov chain yields a closed-form expression for the MTTDL of the system [4]. The results obtained by using CTMC models are often approximate because in practice the times to device failure and the rebuild times are not exponentially distributed. To address this issue, a more general analytical method is required.

B. Non-Markov Analysis

Here, we briefly review the general framework for deriving the MTTDL developed in [5][16] using an analytical approach that does not involve any Markov analysis and therefore avoids the deficiencies of Markov models. The underlying models are not semi-Markov, in that the the system evolution does not depend only on the latest state, but also on the entire path that led to that state. In particular, it depends on the fractions of the data not rebuilt when entering each state. In [21], it was demonstrated that a careless evaluation of these fractions may in fact easily lead to erroneous results.

At any point in time, the system can be thought to be in one of two modes: normal mode and rebuild mode. During normal mode, all data in the system has the original amount of redundancy and there is no active rebuild in process. During rebuild mode, some data in the system has less than the original amount of redundancy and there is an active rebuild process that is trying to restore the redundancy lost. A transition from normal to rebuild mode occurs when a device fails; we refer to the device failure that causes this transition as a *first-device* failure. Following a first-device failure, a complex sequence of rebuild operations and subsequent device failures may occur, which eventually leads the system either to an irrecoverable data loss (DL), with the probability of this event denoted by P_{DL}, or back to the original normal mode by restoring all replicas lost. Typically, the rebuild times are much shorter than the times to failure. Consequently, the time required for this complex sequence of events to complete is negligible compared with the time between successive first-device failures and therefore can be ignored.

Let T_i be the ith interval of a fully operational period, that is, the time interval from the time at which the system is brought to its original state until a subsequent first-device failure occurs. As the system becomes stationary, the length of T_i converges to T. In particular, for a system comprising N devices with a mean time to failure of a device equal to $1/\lambda$, the expected length of T is given by [5]

$$E(T) := \lim_{i\to\infty} E(T_i) = 1/(N\,\lambda)\ . \tag{1}$$

The notation used is given in Table I. Note that the methodology presented here does not involve any Markov analysis and holds for general failure time distributions, which can be exponential or non-exponential, such as the Weibull and gamma distributions.

As the probability that each first-device failure results in data loss is P_{DL}, the expected number of first-device failures until data loss occurs is $1/P_{\text{DL}}$. Thus, by neglecting the effect of the relatively short transient rebuild periods of the system, the MTTDL is essentially the product of the expected time between two first-device-failure events, $E(T)$, and the expected number of first-device-failure events, $1/P_{\text{DL}}$:

$$\text{MTTDL} \approx \frac{E(T)}{P_{\text{DL}}}\ . \tag{2}$$

Substituting (1) into (2) yields

$$\text{MTTDL} \approx \frac{1}{N\,\lambda\,P_{\text{DL}}}\ . \tag{3}$$

III. Direct Path to Data Loss

As mentioned in Section II, during rebuild mode, some data in the system has less than the original amount of redundancy and there is an active rebuild process that aims at restoring the lost redundancy. The direct path to data loss represents the most likely scenario that leads to data loss. This path considers the smallest number of subsequent device failures that occur while the system is in rebuild mode and lead to data loss.

The direct-path-approximation method was applied in [5][16] and led to an analytical approach that does not involve any Markov analysis and therefore avoids the deficiencies of Markov models. This approach yields accurate results when the storage devices are highly reliable, that is, when the ratio of the mean rebuild time $1/\mu$ (typically on the order of tens of hours) to the mean time to failure of a device $1/\lambda$ (typically on the order of a few years) is very small:

$$\frac{1}{\mu} \ll \frac{1}{\lambda}\ , \quad \text{or} \quad \frac{\lambda}{\mu} \ll 1\ , \quad \text{or} \quad \lambda \ll \mu\ . \tag{4}$$

More specifically, this approach considers the system to be in exposure level e when the maximum number of replicas lost by any of the data (or the maximum number of codeword symbols lost in an erasure-coded system) is equal to e. Let us consider, for instance, a replication-based storage system where user data is replicated r times. In this case, the system is in exposure level e if there exists data with $r-e$ copies, but there is no data with fewer than $r-e$ copies. Device failures and rebuild processes cause the exposure level to vary over time. Consider the direct path of successive transitions from exposure level 1 to r. In [16], it was shown that P_{DL} can be approximated by the probability of the direct path to data loss, $P_{\text{DL,direct}}$, when devices are highly reliable, that is,

$$P_{\text{DL}} \approx P_{\text{DL,direct}} = \prod_{e=1}^{r-1} P_{e\to e+1}, \tag{5}$$

where $P_{e\to e+1}$ denotes the transition probability from exposure level e to $e+1$. In fact, the above approximation holds for arbitrary device failure time distributions, and the relative error tends to zero as for highly reliable devices the ratio λ/μ tends to zero [5]. The MTTDL is then obtained by substituting (5) into (3). In [18], the direct-path methodology is extended to more general erasure codes, which include RAID systems.

Note that this analysis can also be applied to assess reliability, in terms of the MTTDL, for systems modeled using a CTMC. For instance, in [6], a RAID-5 system that was modeled using a CTMC was analyzed by both a Markov analysis and an approach similar to the general framework. This fact is used in Section IV to compare the MTTDL of RAID systems obtained using the direct-path approximation in the context of the general framework with the corresponding MTTDL obtained using Markov analysis of CTMCs. This approach is simpler, in that it circumvents the inherent complexity of deriving exact MTTDL expressions using Markov analysis. In Section V, we demonstrate that the direct-path-approximation method can be extended and also applied in the case of multiple shortest paths. We establish this for a system modeled using a CTMC, and conjecture that this should also hold in the case of non-Markovian systems.

Note that this method is in contrast to other methods presented in previous works that associate a probability to each device being in a failed state [22]. In particular, those works assume that these probabilities are given and therefore do not account for the rebuild processes, whereas the methods presented in this work do account for the rebuild processes through the probabilities of traversing various states until data loss occurs.

TABLE I. Notation of System Parameters

Parameter	Definition
N	Number of devices in the system
$1/\lambda$	Mean time to failure for a device
$1/\mu$	Mean time to rebuild device failures
$se^{(\text{RAID})}$	Storage efficiency of a RAID scheme
S	Sector size
C_d	Device capacity
P_s	Probability of an unrecoverable or latent sector error
n_s	Number of data sectors in a device ($n_s = C_d/S$)

IV. Comparison of Markov Analysis and Direct-Path Approximation

A common scheme used for tolerating device (disk) failures is the redundant array of independent disks (RAID) [2][3]. The RAID-5 scheme arranges devices in groups (arrays), each with one redundant device, and can tolerate one device failure per array. Similarly, the RAID-6 scheme arranges devices in arrays, each with two redundant devices, and can tolerate up to two device failures per array. Considering a RAID array comprised of N devices, the storage efficiency of a RAID-5 system is given by

$$se^{(\text{RAID-5})} \ = \ \frac{N-1}{N}\ , \tag{6}$$

and the storage efficiency of a RAID-6 system is given by

$$se^{(\text{RAID-6})} \; = \; \frac{N-2}{N} \; . \tag{7}$$

It turns out that the MTTDL of systems comprised of highly reliable devices can be approximated by using the *direct-path approximation*. We proceed to demonstrate this by presenting two specific examples, the RAID-5 and RAID-6 systems. In both cases, the RAID array is assumed to contain N devices, and the numbered states of the corresponding Markov models represent the number of failed devices. The DL state represents a data loss due to a device failure that occurs when the system is in the critical mode of operation. A RAID array is considered to be in *critical mode* when an additional device failure can no longer be tolerated. Thus, RAID-5 and RAID-6 arrays are in critical mode when there are $N-1$ devices and $N-2$ devices in operation, that is, when they operate with one device and two devices failed, respectively.

A. MTTDL *of a RAID-5 Array*

The Markov chain model for a RAID-5 array is shown in Fig. 1. When the first device fails, the array enters critical mode, which corresponds to the transition from state 0 to state 1. As initially there are N devices in operation, the mean time until the first failure is equal to $1/(N\lambda)$, and the corresponding transition rate is its inverse, that is, $N\lambda$. Subsequently, the critical mode ends owing to either a successful completion of the rebuild or another device failure. The former event is represented by the state transition from state 1 to state 0 with a rate of μ, given that the mean rebuild time is equal to $1/\mu$. The latter event leads to data loss and is represented by the state transition from state 1 to state DL with a rate of $(N-1)\lambda$ given that in critical mode there are $N-1$ devices in operation.

The exact MTTDL, denoted by $\text{MTTDL}_{\text{RAID-5}}$, is obtained from [6, Eq. (45)] by setting $P_{\text{uf}}^{(1)} = 0$:

$$\text{MTTDL}_{\text{RAID-5}} \; = \; \frac{\mu + (2N-1)\lambda}{N\,(N-1)\,\lambda^2} \; . \tag{8}$$

Note that when $\lambda \ll \mu$, the first term of the numerator in (8) can be ignored, such that the $\text{MTTDL}_{\text{RAID-5}}$ can be approximated by $\text{MTTDL}_{\text{RAID-5}}^{(\text{approx})}$ as follows:

$$\text{MTTDL}_{\text{RAID-5}}^{(\text{approx})} \; \approx \; \frac{\mu}{N(N-1)\,\lambda^2} \; . \tag{9}$$

This result was obtained in [2] by using an approach that is essentially the direct-path approximation. Next, we present its derivation for completeness. The transition from state 0 to state 1 represents the first device failure. The direct path to data loss involves a subsequent device failure before it can complete the rebuild process and return to state 0. This corresponds to the state transition from state 1 to state DL, with the corresponding probability $P_{1\to\text{DL}}$ given by

$$P_{\text{DL}} \approx P_{\text{DL,direct}} = P_{1\to\text{DL}} = \frac{(N-1)\,\lambda}{\mu + (N-1)\,\lambda} \tag{10}$$

$$\approx (N-1)\left(\frac{\lambda}{\mu}\right) \; , \tag{11}$$

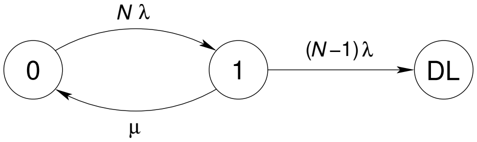

Figure 1. Reliability model for a RAID-5 array.

where the approximation is obtained by using (4) and therefore neglecting the second term of the denominators in (10). Substituting (10) into (3) yields

$$\text{MTTDL}'_{\text{RAID-5}} \; \approx \; \frac{\mu + (N-1)\,\lambda}{N(N-1)\,\lambda^2} \; . \tag{12}$$

Note that the approximation given in (9) now follows immediately from (12) by using (4) and therefore neglecting the second term of the numerator.

B. MTTDL *of a RAID-6 Array*

The Markov chain model for a RAID-6 array is shown in Fig. 2. The first device failure is represented by the transition from state 0 to state 1. As initially there are N devices in operation, the mean time until the first failure is $1/(N\lambda)$, and the corresponding transition rate is its inverse, that is, $N\lambda$. The system exits from state 1 owing to either a successful completion of the rebuild or another device failure. The former event is represented by the state transition from state 1 to state 0 with a rate of μ. The latter event is represented by the state transition from state 1 to state 2 with a rate of $(N-1)\lambda$. Subsequently, the system exits from state 2 owing to either a successful completion of the rebuild or another device failure. The former event is represented by the state transition from state 2 to state 0 with a rate of μ, given that the mean rebuild time is equal to $1/\mu$. The latter event leads to data loss and is represented by the state transition from state 2 to state DL with a rate of $(N-2)\lambda$ given that in critical mode there are $N-2$ devices in operation.

The exact MTTDL, denoted by $\text{MTTDL}_{\text{RAID-6}}$, is obtained from [6, Eq. (52)] by setting $\mu_1 = \mu_2 = \mu$ and $P_{\text{uf}}^{(\text{r})} = P_{\text{uf}}^{(2)} = 0$:

$$\text{MTTDL}_{\text{RAID-6}} \; = \; \frac{\mu^2 + 3(N-1)\lambda\mu + (3N^2 - 6N + 2)\,\lambda^2}{N\,(N-1)\,(N-2)\,\lambda^3} \; . \tag{13}$$

Note that when $\lambda \ll \mu$, the last two terms of the numerator of (13) can be neglected and thus $\text{MTTDL}_{\text{RAID-6}}$ can be approximated by $\text{MTTDL}_{\text{RAID-6}}^{(\text{approx})}$ as follows:

$$\text{MTTDL}_{\text{RAID-6}}^{(\text{approx})} \; \approx \; \frac{\mu^2}{N(N-1)(N-2)\,\lambda^3} \; , \tag{14}$$

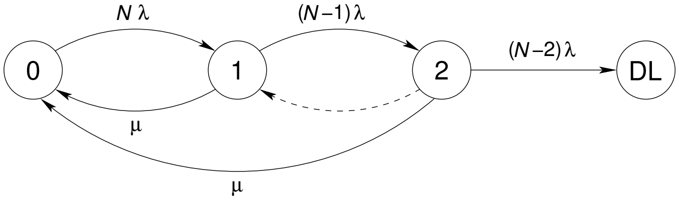

Figure 2. Reliability model for a RAID-6 array.

which is the same result as that reported in [3].

We now proceed to show how the approximate MTTDL of the system can be derived in a straightforward manner by applying the direct-path-approximation technique. The transition from state 0 to state 1 represents the first device failure. The direct path to data loss involves two subsequent device failures before it can complete the rebuild process and return to state 0. This corresponds to the state transitions from state 1 to state 2 and from state 2 to state DL, with the corresponding probabilities $P_{1\to 2}$ and $P_{2\to \mathrm{DL}}$ given by

$$P_{1\to 2} = \frac{(N-1)\,\lambda}{\mu + (N-1)\,\lambda} \; . \tag{15}$$

and

$$P_{2\to \mathrm{DL}} = \frac{(N-2)\,\lambda}{\mu + (N-2)\,\lambda} \; . \tag{16}$$

Thus, the probability of data loss, that is, the probability that from state 1 the system goes to state DL before it can reach state 0, is equal to

$$\begin{aligned} P_{\mathrm{DL}} \approx P_{\mathrm{DL,direct}} &= P_{1\to 2}\, P_{2\to \mathrm{DL}} \\ &= \frac{(N-1)\,\lambda}{\mu + (N-1)\,\lambda} \cdot \frac{(N-2)\,\lambda}{\mu + (N-2)\,\lambda} \qquad (17) \\ &\approx (N-1)(N-2)\left(\frac{\lambda}{\mu}\right)^2 \; , \qquad (18) \end{aligned}$$

where the approximation is obtained by using (4) and therefore neglecting the second terms of the denominators in (17).

We verify that substituting (18) into (3) yields the approximation given in (14).

Remark 1: If the transition from state 2 to state 0 were not to state 0 but to state 1 instead, as shown in Fig. 2 by the dashed arrow, the expression for $P_{2\to \mathrm{DL}}$ given by (16) would still hold. However, in this case it would hold that $P_{\mathrm{DL}} > P_{\mathrm{DL,direct}}$ as, in addition to the direct path $1 \to 2 \to \mathrm{DL}$, there are other possible paths $1 \to 2 \to 1 \to 2 \to \cdots \to 1 \to 2 \to \mathrm{DL}$ to data loss. In [16] it was shown that, for systems with highly reliable components, the direct path dominates the effect of all other possible paths and therefore its probability, $P_{\mathrm{DL,direct}}$, approximates well the probability of all paths, P_{DL}, that is,

$$P_{\mathrm{DL}} \approx P_{\mathrm{DL,direct}} = P_{1\to 2}\, P_{2\to \mathrm{DL}} \approx \frac{(N-1)(N-2)\,\lambda^2}{\mu^2} \; . \tag{19}$$

In this case, the MTTDL is given by

$$\mathrm{MTTDL}'_{\text{RAID-6}} \;=\; \frac{(3N^2 - 6N + 2)\,\lambda^2 + 2(N-1)\lambda\mu + \mu^2}{N\,(N-1)\,(N-2)\,\lambda^3} \; , \tag{20}$$

which, as expected, is less than that given in (13). Despite this difference, the approximation given in (14) still holds because (19) is the same as (18).

V. Multiple Shortest Paths to Data Loss

Here, we consider redundancy schemes for which there are multiple shortest paths to data loss. Following the analysis presented in [16] for the direct-path approximation, we conjecture that, for systems with highly reliable devices, the shortest paths dominate the effect of all other possible paths and therefore the sum of their corresponding probabilities, $P_{\mathrm{DL,shortest}}$, approximates well the probability of all paths, P_{DL}, that is,

$$P_{\mathrm{DL}} \approx P_{\mathrm{DL,shortest}} \; . \tag{21}$$

A. A RAID-51 Array

We proceed by considering a RAID-51 system, which is a RAID-5 array with mirroring. The contents of failed devices are recovered by their mirrors, and if this is not possible, they are recovered through the corresponding RAID-5 arrays. The configuration comprises D pairs of mirrored devices, where each pair contains two devices with identical content. Therefore, it consists of two identical RAID-5 arrays, for a total of $N = 2\,D$ devices and a storage efficiency given by

$$se^{(\text{RAID-51})} \;=\; \frac{D-1}{2D} \;=\; \frac{N-2}{2N} \; . \tag{22}$$

This configuration was considered in [11], referred to as RAID 5+1, with the corresponding Markov model shown in [11, Fig. 7(a)]. It is redrawn in Fig. 3 with the parameters λ and μ corresponding to the parameters μ and ν of the initial figure, respectively. Also, the DL states correspond to the 'Failure' states, and the state tuples (x, y, z) indicate that there are x pairs with both devices in operation, y pairs with one device in operation and one device failed, and z pairs with both devices failed. Also, some typos regarding the transition rates were corrected.

An exact evaluation of the MTTDL associated with this Markov chain model appears to be a very challenging, if not infeasible, task. Thus, in [11] a rough approximation was obtained by first deriving the failure and repair rates for a mirrored pair of devices, and then substituting these values into expression (9) for a single RAID-5 system. The MTTDL is obtained in [11, Eq. (11)] as follows:

$$\mathrm{MTTDL} \;\approx\; \frac{\mu^3}{4D(D-1)\,\lambda^4} \; . \tag{23}$$

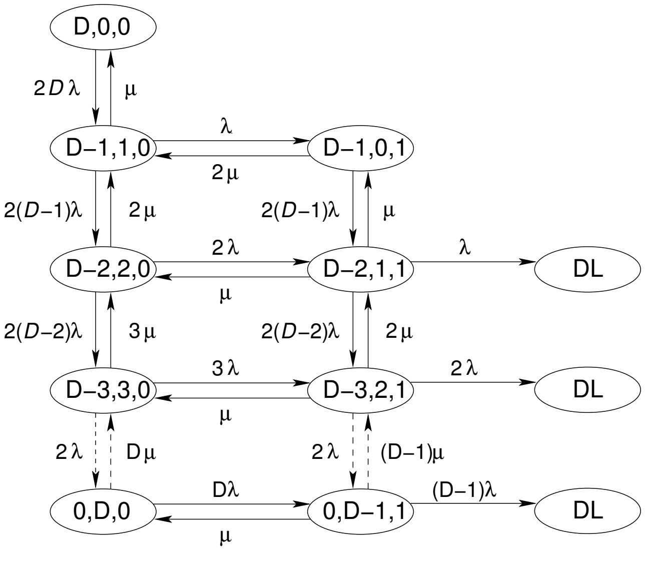

Figure 3. Reliability model for a RAID-51 array.

1) MTTDL *Evaluation Using the Shortest-Path Approximation:* The transition from state $(D, 0, 0)$ to state $(D-1, 1, 0)$ represents the first device failure. As initially there are $2\,D$ devices in operation, the mean time until the first failure is $1/(2D\lambda)$, and the corresponding transition rate is its inverse, $2D\lambda$.

The most likely path to data loss is the shortest path from state $(D-1, 1, 0)$ to a DL state, which in this case comprises two such paths, as shown in Fig. 4: the upper path $(D-1, 1, 0) \rightarrow (D-1, 0, 1) \rightarrow (D-2, 1, 1) \rightarrow$ DL and the lower path $(D-1, 1, 0) \rightarrow (D-2, 2, 0) \rightarrow (D-2, 1, 1) \rightarrow$ DL. Each of these paths involves three subsequent device failures.

After the first device has failed, there are $D-1$ pairs with both devices in operation, and one pair, say PR_1, with one device in operation and one device failed, which corresponds to the transition from state $(D, 0, 0)$ to state $(D-1, 1, 0)$. The rebuild of the failed device consists of recovering its data to a new spare device by copying the contents of its mirror to it, that is, of the device in operation in PR_1. Then, the next event can be either a successful completion of the rebuild or another device failure. The former event is represented by the state transition from state $(D-1, 1, 0)$ to state $(D, 0, 0)$ with a rate of μ. For the latter event, two cases are considered:

Case 1: Upper path. The second device that fails is the device in operation concerning pair PR_1, which corresponds to the transition from state $(D-1, 1, 0)$ to state $(D-1, 0, 1)$, as now both devices of pair PR_1 have failed, and all other $D-1$ pairs remain intact. The transition rate is λ, which is the failure rate of the last failed device. The contents of the devices of pair PR_1 are recovered through the corresponding RAID-5 arrays. As both devices of pair PR_1 are under rebuild, the transition rate from state $(D-1, 0, 1)$ back to state $(D-1, 1, 0)$ is 2μ. If, however, prior to the completion of any of these two rebuilds another device of the remaining $2(D-1)$ devices fails, then there will be $D-2$ pairs with both devices in operation, one pair, say PR_2, with one device in operation and one device failed, and pair PR_1 with both devices failed. This corresponds to the transition from state $(D-1, 0, 1)$ to state $(D-2, 1, 1)$, with a transition rate equal to $2(D-1)\lambda$. Note that in [11, Fig. 7(a)] this transition rate is erroneously indicated as $(2D-1)\mu$ instead of $2(D-1)\mu$.

Case 2: Lower path. The second device that fails is one of the $2(D-1)$ devices in the $D-1$ pairs, say a device concerning PR_2. This corresponds to the transition from state $(D-1, 1, 0)$ to state $(D-2, 2, 0)$, as both pairs PR_1 and PR_2 now have one device in operation and one device failed, and all other $D-2$ pairs remain intact. The corresponding transition rate is equal to $2(D-1)\lambda$. Note that in [11, Fig. 7(a)] this transition rate is erroneously indicated as $(2D-1)\mu$ instead of $2(D-1)\mu$. The contents of the failed devices are recovered from their corresponding mirrors. As both devices of the two pairs PR_1 and PR_2 are under rebuild, the transition rate from state $(D-2, 2, 0)$ back to state $(D-1, 1, 0)$ is 2μ. If, however, prior to the completion of any of these two rebuilds another device of the two remaining devices in operation in PR_1 and PR_2 fails (say, that of pair PR_1), then there will be $D-2$ pairs with both devices in operation, one pair (PR_2) with one device in operation and one device failed, and one pair (PR_1) with both devices failed. This corresponds to the transition from state $(D-2, 2, 0)$ to state $(D-2, 1, 1)$, with a transition rate 2λ.

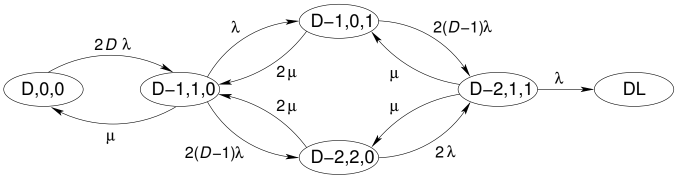

Figure 4. Shortest-path reliability model for a RAID-51 array.

At state $(D-2, 1, 1)$, the failed device in pair PR_2 is recovered by its mirror. However, the corresponding failed device in pair PR_1 cannot be recovered because the RAID-5 array has suffered two device failures. In contrast, the failed device in pair PR_1 can be recovered because the corresponding RAID-5 array has suffered only one device failure.

The completion of the rebuild of the failed device in pair PR_2 corresponds to the transition from state $(D-2, 1, 1)$ to state $(D-1, 0, 1)$, with a transition rate of μ. The completion of the rebuild of the failed device in pair PR_1 through the RAID capability corresponds to the transition from state $(D-2, 1, 1)$ to state $(D-2, 2, 0)$, with a transition rate of μ. Note that in [11, Fig. 7(a)] this transition rate is erroneously indicated as 2μ instead of μ. If, however, prior to the completion of any of these rebuilds, the device still in operation of pair PR_2 fails, this leads to data loss, as there will be two pairs failed, with each of the RAID-5 arrays having two devices failed. This corresponds to the transition from state $(D-2, 1, 1)$ to state DL, with a corresponding rate of λ.

The probabilities of the transitions discussed above are given by

$$P_{(D-1,1,0)\rightarrow(D-1,0,1)} = \frac{\lambda}{\mu + (2D-1)\,\lambda}\,, \tag{24}$$

$$P_{(D-1,0,1)\rightarrow(D-2,1,1)} = \frac{2(D-1)\,\lambda}{2\,\mu + 2(D-1)\,\lambda}\,, \tag{25}$$

$$P_{(D-1,1,0)\rightarrow(D-2,2,0)} = \frac{2(D-1)\,\lambda}{\mu + (2D-1)\,\lambda}\,, \tag{26}$$

$$P_{(D-2,2,0)\rightarrow(D-2,1,1)} = \frac{2\,\lambda}{2\,\mu + 2\,\lambda}\,, \tag{27}$$

and

$$P_{(D-2,1,1)\rightarrow\mathrm{DL}} = \frac{\lambda}{2\,\mu + \lambda}\,. \tag{28}$$

Consequently, the probability of the upper path to data loss, P_u, is given by

$$\begin{aligned} P_u &= P_{(D-1,1,0)\rightarrow(D-1,0,1)}\,P_{(D-1,0,1)\rightarrow(D-2,1,1)}\,P_{(D-2,1,1)\rightarrow\mathrm{DL}} \\ &= \frac{\lambda}{\mu + (2D-1)\,\lambda} \cdot \frac{2(D-1)\,\lambda}{2\,\mu + 2(D-1)\,\lambda} \cdot \frac{\lambda}{2\,\mu + \lambda}\,, \end{aligned} \tag{29}$$

and that of the lower path to data loss, P_l, is given by

$$\begin{aligned} P_l &= P_{(D-1,1,0)\rightarrow(D-2,2,0)}\,P_{(D-2,2,0)\rightarrow(D-2,1,1)}\,P_{(D-2,1,1)\rightarrow\mathrm{DL}} \\ &= \frac{2(D-1)\,\lambda}{\mu + (2D-1)\,\lambda} \cdot \frac{2\,\lambda}{2\,\mu + 2\,\lambda} \cdot \frac{\lambda}{2\,\mu + \lambda}\,. \end{aligned} \tag{30}$$

By considering (4), (29) and (30) yield the following approximations:

$$P_u \approx \frac{\lambda}{\mu} \cdot \frac{2(D-1)\,\lambda}{2\,\mu} \cdot \frac{\lambda}{2\,\mu} = \frac{(D-1)\,\lambda^3}{2\mu^3} \qquad (31)$$

and

$$P_l \approx \frac{2(D-1)\,\lambda}{\mu} \cdot \frac{\lambda}{\mu} \cdot \frac{\lambda}{2\,\mu} = \frac{(D-1)\,\lambda^3}{\mu^3} \,. \qquad (32)$$

The probability of the shortest paths to data loss, $P_{\text{DL,shortest}}$, is the sum of P_u and P_l, which by using (21), (31), and (32), yields

$$P_{\text{DL}} \approx P_{\text{DL,shortest}} = P_u + P_l \approx \frac{3(D-1)}{2} \left(\frac{\lambda}{\mu}\right)^3 \,. \qquad (33)$$

Substituting (33) into (3), and considering $N = 2D$, yields the approximate MTTDL of the RAID-51 system, $\text{MTTDL}_{\text{RAID-51}}^{\text{(approx)}}$, given by

$$\text{MTTDL}_{\text{RAID-51}}^{\text{(approx)}} \approx \frac{\mu^3}{3D(D-1)\,\lambda^4} \,. \qquad (34)$$

Remark 2: Note that the prediction given by (34) is higher than that obtained in [11], which is given by (23). At first glance, this seems to be counterintuitive. The approximation in [11] considers only failures of mirrored device pairs, which corresponds to the upper path to data loss. As this neglects the lower path, one would expect the prediction in [11] to be higher, not lower. The reason for this counterintuitive result is that considering additional paths may, on the one hand increase the number of paths that lead to data loss, but on the other hand it may also increase the number of paths that do not lead to data loss, therefore delaying the occurrence of data loss. For instance, when the lower path is neglected, the probability $P_{(D-2,1,1)\to\text{DL}}$ of the transition from state $(D-2,1,1)$ to state DL is equal to $\lambda/(\lambda+\mu)$, which is greater than the corresponding one given by (28) if also the lower path is considered.

2) Exact MTTDL *Evaluation for D = 3:* An exact evaluation of the reliability of a RAID-51 system through the MTTDL associated with the corresponding Markov chain model shown in Fig. 3 appears to be a very challenging, if not infeasible, task for arbitrary D. Therefore, we proceed by considering a RAID-51 system with $D = 3$. The corresponding Markov chain model is shown in Fig. 5. The exact MTTDL of this system, denoted by $\text{MTTDL}_{\text{RAID-51}}^{(D=3)}$, is obtained by using the infinitesimal generator matrix approach and determining the average time spent in the transient states of the Markov chain [4]. Because of space limitations, we only provide the final result:

$$\text{MTTDL}_{\text{RAID-51}}^{(D=3)} = \frac{2+20\frac{\lambda}{\mu}+93(\frac{\lambda}{\mu})^2+287(\frac{\lambda}{\mu})^3+677(\frac{\lambda}{\mu})^4+939(\frac{\lambda}{\mu})^5+630(\frac{\lambda}{\mu})^6}{12\,\lambda^4\,\mu^{-3}\,[3+18\frac{\lambda}{\mu}+35(\frac{\lambda}{\mu})^2+30(\frac{\lambda}{\mu})^3]} \,. \qquad (35)$$

Note that when $\lambda \ll \mu$, $\text{MTTDL}_{\text{RAID-51}}^{(D=3)}$ can be approximated by $\text{MTTDL}_{\text{RAID-51}}^{(D=3,\text{approx})}$ as follows:

$$\text{MTTDL}_{\text{RAID-51}}^{(D=3,\text{approx})} \approx \frac{\mu^3}{18\,\lambda^4} \,, \qquad (36)$$

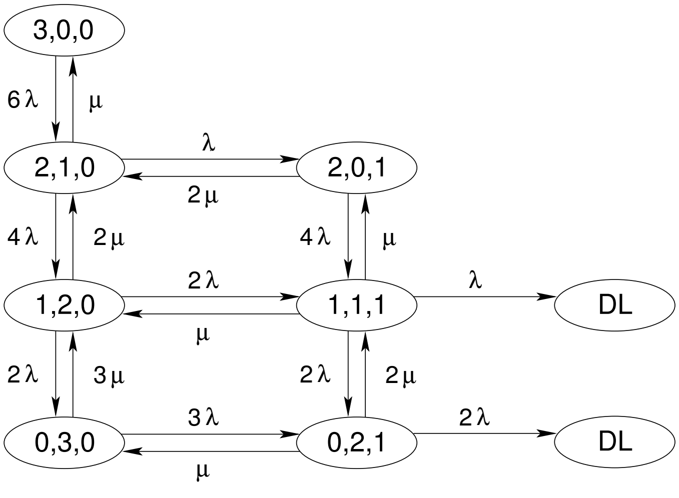

Figure 5. Reliability model for a RAID-51 array with $D = 3$.

which is the same result as that predicted by (34) for $D = 3$ and therefore confirms its validity.

B. A Two-Dimensional RAID-5 Array

We consider a two-dimensional RAID-5 array with the devices arranged in K rows and D columns for a total of $N = KD$ devices, including superparity [13]. In this configuration, denoted by 2D-RAID-5, the devices in each row and each column form a RAID-5 array with the corresponding storage efficiency given by

$$se^{(\text{2D-RAID-5})} = \frac{(K-1)\,(D-1)}{K\,D} \,. \qquad (37)$$

The contents of failed devices are recovered either horizontally or vertically through the corresponding RAID-5 arrays. This system tolerates all triple device failures and can also tolerate more than three device failures for certain constellations, e.g., the failure of an entire row or column. However, as the devices are assumed to fail independently, the shortest path to data loss is due to the failure of four devices occurring in the constellation shown in Fig. 6 with the failed devices located in two rows and two columns. The special case of a specific square RAID-5 array (i.e., $K = D$) was considered in [20], but no closed-form expression for the MTTDL was provided owing to its complexity. Here, we obtain approximate closed-form expressions for the MTTDL that are general, simple, yet accurate for real-world systems. Fig. 6 also shows the Markov chain model corresponding to the shortest path to data loss. The state tuples (x,y,z,w) indicate that there are x rows with one device failed and $D-1$ devices in operation, y rows with two devices failed and $D-2$ devices in operation, z columns with one device failed and $N-1$ devices in operation, and w columns with two devices failed and $N-2$ devices in operation. The relevant states are shown next to the corresponding device failure constellations indicated with ‘x’ on the $K \times D$ plane.

We now proceed to evaluate the MTTDL using the shortest-path approximation. The transition from state $(0,0,0,0)$ to state $(1,0,1,0)$ represents the first device failure. The

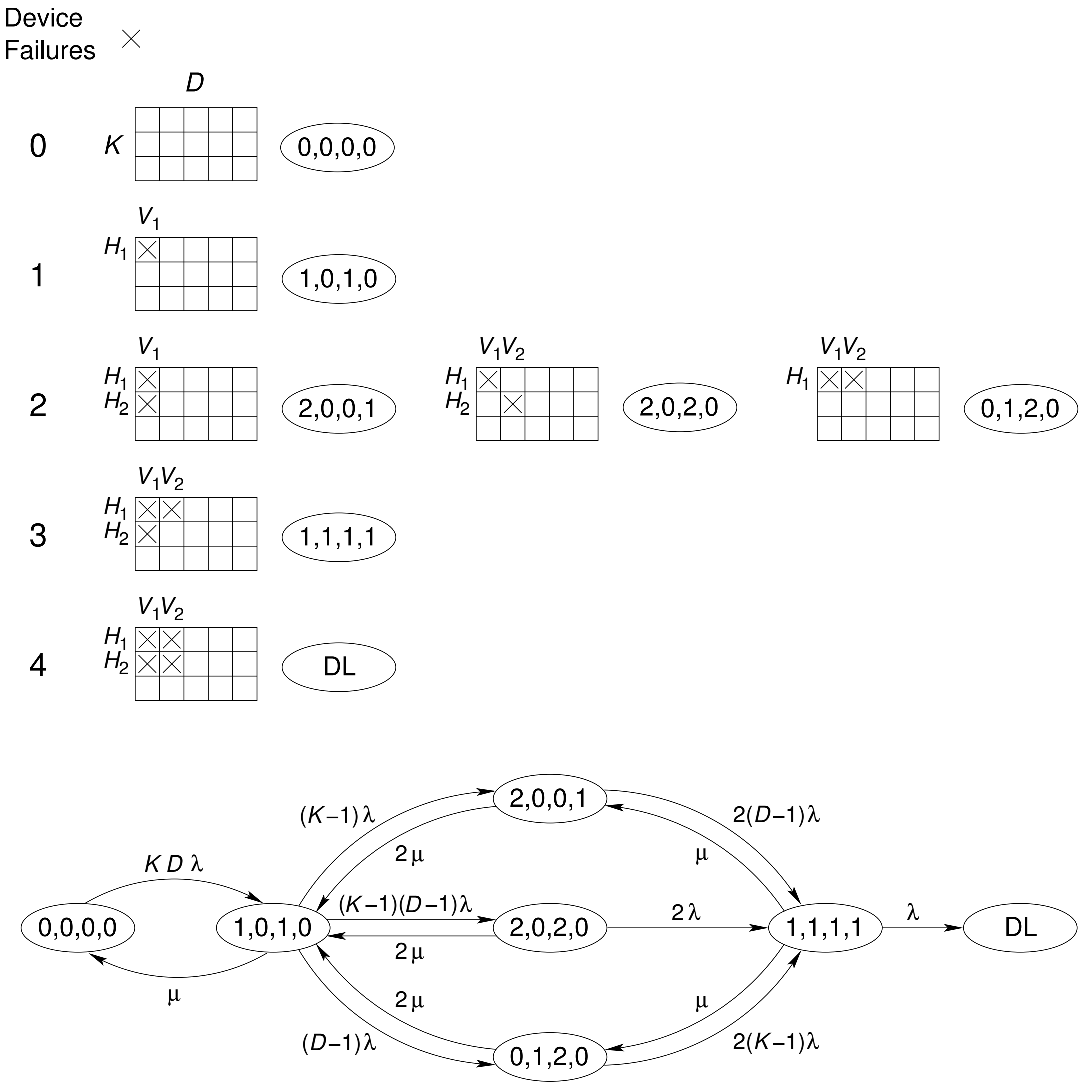

Figure 6. Shortest-path reliability model for a two-dimensional RAID-5 array.

most likely path to data loss is the shortest path from state $(1, 0, 1, 0)$ to the DL state, which in this case comprises three such paths, as shown in Fig. 6: the upper path $(1, 0, 1, 0) \rightarrow (2, 0, 0, 1) \rightarrow (1, 1, 1, 1) \rightarrow$ DL, the middle path $(1, 0, 1, 0) \rightarrow (2, 0, 2, 0) \rightarrow (1, 1, 1, 1) \rightarrow$ DL, and the lower path $(1, 0, 1, 0) \rightarrow (0, 1, 2, 0) \rightarrow (1, 1, 1, 1) \rightarrow$ DL. Each of these paths involves three subsequent device failures.

After the first device has failed, there are two RAID-5 arrays, one horizontal RAID-5 array, say row H_1, and one vertical RAID-5 array, say column V_1, with one device failed in each of them. Consequently, this failure corresponds to the transition from state $(0, 0, 0, 0)$ to state $(1, 0, 1, 0)$. As initially there are KD devices in operation, the mean time until the first failure is $1/(KD\lambda)$, and the corresponding transition rate is its inverse, $KD\lambda$. The rebuild of the failed device can be performed using either $H1$ or $V1$. Then, the next event can be either a successful completion of the rebuild or another device failure. The former event is represented by the state transition from state $(1, 0, 1, 0)$ to state $(0, 0, 0, 0)$, with a rate of μ. For the latter event, three cases are considered:

Case 1: Upper path. The second device that fails is one of the $K-1$ operating devices in V_1. This occurs with a transition rate of $(K-1)\lambda$ and results in another horizontal RAID-5 array, say, row H_2, with one device failed. This corresponds to the transition from state $(1, 0, 1, 0)$ to state $(2, 0, 0, 1)$, as there are now two rows with one device failed in each of them and one column with two devices failed. As the contents of the two failed devices in V_1 are rebuilt in parallel through the corresponding H_1 and H_2 RAID-5 arrays, the transition rate from state $(2, 0, 0, 1)$ back to state $(1, 0, 1, 0)$ is 2μ. If, however, prior to the completion of any of these two rebuilds, another of the remaining $2(D-1)$ devices in H_1 and H_2 (say, the one in row H_1 and column V_2) fails, then there will be
– a column, namely V_2, with one device failed;
– a row, namely H_1, with two devices failed;
– a row, namely H_2, with one device failed, and
– a column, namely V_1, with two devices failed.
Note that the $(K-2)D$ operational devices in the remaining $K-2$ horizontal RAID-5 arrays are not considered because their failure leads to states that are not in the shortest paths. The above corresponds to the transition from state $(2, 0, 0, 1)$

to state $(1,1,1,1)$, with a transition rate equal to $2(D-1)\lambda$.

Case 2: Middle path. The second device that fails is one of the $(K-1)(D-1)$ operating devices that are not in H_1 or V_1. This occurs with a transition rate of $(K-1)(D-1)\lambda$ and results in another horizontal RAID-5 array, say, row H_2, and another vertical RAID-5 array, say, column V_2, with one device failed. This corresponds to the transition from state $(1,0,1,0)$ to state $(2,0,2,0)$ as there are now two rows and two columns with one device failed in each of them. As the contents of the two failed devices are rebuilt in parallel through the corresponding, horizontal or vertical, RAID-5 arrays, the transition rate from state $(2,0,0,1)$ back to state $(1,0,1,0)$ is 2μ. If, however, prior to the completion of any of these two rebuilds, either the (H_1, V_2) or the (H_2, V_1) device fails, where (H_1, V_2) refers to the device in row H_1 and column V_2, and (H_2, V_1) to that in row H_2 and column V_1, then there will be
– a column and a row with one device failed, and
– a column and a row with two devices failed.
Note that the remaining $KD-4$ operational devices are not considered because their failure leads to states that are not in the shortest paths. The above corresponds to the transition from state $(2,0,2,0)$ to state $(1,1,1,1)$, with a transition rate equal to 2λ.

Case 3: Lower path. The second device that fails is one of the $D-1$ operating devices in H_1. This occurs with a transition rate of $(D-1)\lambda$ and results in another vertical RAID-5 array, say column V_2, with one device failed. This corresponds to the transition from state $(1,0,1,0)$ to state $(0,1,2,0)$, as there are now one row with two devices failed and two columns with one device failed in each of them. As the contents of the two failed devices in H_1 are rebuilt in parallel through the corresponding V_1 and V_2 RAID-5 arrays, the transition rate from state $(2,0,0,1)$ back to state $(1,0,1,0)$ is 2μ. If, however, prior to the completion of any of these two rebuilds another of the remaining $2(K-1)$ devices in V_1 and V_2 (say the one in row H_2 and column V_1) fails, then there will be
– a column, namely V_2, with one device failed;
– a row, namely H_1, with two devices failed;
– a row, namely H_2, with one device failed, and
– a column, namely V_1, with two devices failed.
Note that the $(D-2)K$ operational devices in the remaining $D-2$ vertical RAID-5 arrays are not considered because their failure leads to states that are not in the shortest paths. The above corresponds to the transition from state $(0,1,2,0)$ to state $(1,1,1,1)$, with a transition rate equal to $2(K-1)\lambda$.

At state $(1,1,1,1)$, the failed device in H_2 and the failed one in V_2 are recovered through their corresponding RAID-5 arrays. However, the failed device in row H_1 and column V_1 cannot be immediately recovered because both of its corresponding RAID-5 arrays has suffered two device failures. It can only be recovered upon completion of the rebuild of either one of the two previously mentioned devices. In particular, the completion of the rebuild of the failed device in V_2 corresponds to the transition from state $(1,1,1,1)$ to state $(2,0,0,1)$, with a transition rate of μ. The completion of the rebuild of the failed device in H_2 corresponds to the transition from state $(1,1,1,1)$ to state $(0,1,2,0)$, with a transition rate of μ. If, however, prior to the completion of any of these two rebuilds, the device still in operation in row H_2 and column V_2 fails, this leads to data loss, as there will be four failed devices with each of the corresponding RAID-5 arrays having two failed devices. This corresponds to the transition from state $(1,1,1,1)$ to state DL, with a corresponding rate of λ.

The probabilities of the transitions discussed above are given by

$$P_{(1,0,1,0)\to(2,0,0,1)} = \frac{(K-1)\,\lambda}{\mu+(KD-1)\,\lambda}\,, \tag{38}$$

$$P_{(2,0,0,1)\to(1,1,1,1)} = \frac{2(D-1)\,\lambda}{2\,\mu+2(D-1)\,\lambda}\,, \tag{39}$$

$$P_{(1,0,1,0)\to(2,0,2,0)} = \frac{(K-1)(D-1)\,\lambda}{\mu+(KD-1)\,\lambda}\,, \tag{40}$$

$$P_{(2,0,2,0)\to(1,1,1,1)} = \frac{2\,\lambda}{2\,\mu+2\,\lambda}\,, \tag{41}$$

$$P_{(1,0,1,0)\to(0,1,2,0)} = \frac{(D-1)\,\lambda}{\mu+(KD-1)\,\lambda}\,, \tag{42}$$

$$P_{(0,1,2,0)\to(1,1,1,1)} = \frac{2(K-1)\,\lambda}{2\,\mu+2(K-1)\,\lambda}\,, \tag{43}$$

and

$$P_{(1,1,1,1)\to\mathrm{DL}} = \frac{\lambda}{2\,\mu+\lambda}\,. \tag{44}$$

Consequently, the probability of the upper path to data loss, P_u, is given by

$$\begin{aligned} P_u &= P_{(1,0,1,0)\to(2,0,0,1)}\, P_{(2,0,0,1)\to(1,1,1,1)}\, P_{(1,1,1,1)\to\mathrm{DL}} \\ &= \frac{(K-1)\lambda}{\mu+(KD-1)\,\lambda}\cdot\frac{2(D-1)\,\lambda}{2\,\mu+2(D-1)\,\lambda}\cdot\frac{\lambda}{2\,\mu+\lambda}\,, \end{aligned} \tag{45}$$

that of the middle path to data loss, P_m, is given by

$$\begin{aligned} P_m &= P_{(1,0,1,0)\to(2,0,2,0)}\, P_{(2,0,2,0)\to(1,1,1,1)}\, P_{(1,1,1,1)\to\mathrm{DL}} \\ &= \frac{(K-1)(D-1)\lambda}{\mu+(KD-1)\,\lambda}\cdot\frac{2\,\lambda}{2\,\mu+2\,\lambda}\cdot\frac{\lambda}{2\,\mu+\lambda}\,, \end{aligned} \tag{46}$$

and that of the lower path to data loss, P_l, is given by

$$\begin{aligned} P_l &= P_{(1,0,1,0)\to(0,1,2,0)}\, P_{(0,1,2,0)\to(1,1,1,1)}\, P_{(1,1,1,1)\to\mathrm{DL}} \\ &= \frac{(D-1)\,\lambda}{\mu+(KD-1)\,\lambda}\cdot\frac{2(K-1)\,\lambda}{2\,\mu+2(K-1)\,\lambda}\cdot\frac{\lambda}{2\,\mu+\lambda}\,. \end{aligned} \tag{47}$$

By considering (4), equations (45), (46), and (47) yield the following approximations:

$$P_u \;\approx\; \frac{(K-1)\lambda}{\mu}\cdot\frac{2(D-1)\,\lambda}{2\,\mu}\cdot\frac{\lambda}{2\,\mu} \;=\; \frac{(K-1)(D-1)\,\lambda^3}{2\mu^3}\,, \tag{48}$$

$$P_m \;\approx\; \frac{(K-1)(D-1)\lambda}{\mu}\cdot\frac{2\,\lambda}{2\,\mu}\cdot\frac{\lambda}{2\,\mu} \;=\; \frac{(K-1)(D-1)\,\lambda^3}{2\mu^3}\,, \tag{49}$$

and

$$P_l \;\approx\; \frac{(D-1)\,\lambda}{\mu}\cdot\frac{2(K-1)\,\lambda}{2\,\mu}\cdot\frac{\lambda}{2\,\mu} \;=\; \frac{(K-1)(D-1)\,\lambda^3}{2\mu^3}\,. \tag{50}$$

The probability of the shortest paths to data loss, $P_{\text{DL,shortest}}$, is the sum of P_u, P_m and P_l, which by using (21), (48), (49), and (50), yields

$$P_{\text{DL}} \approx P_{\text{DL,shortest}} = P_u + P_m + P_l \approx \frac{3(K-1)(D-1)}{2}\left(\frac{\lambda}{\mu}\right)^3 . \tag{51}$$

Substituting (51) into (3), and considering $N = KD$, yields the approximate MTTDL of the two-dimensional RAID-5 system, $\text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}}$, given by

$$\text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \approx \frac{2\,\mu^3}{3\,K\,(K-1)\,D\,(D-1)\,\lambda^4} . \tag{52}$$

VI. Reliability Comparison

Here, we assess the relative reliability of the various schemes considered. As discussed in Section III, the direct-path-approximation method yields accurate results 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 perform a fair comparison by considering systems with the same amount of user data stored under the same storage efficiency. Note that, according to (6) and (22), the storage efficiency of a RAID-5 system cannot be less than 1/2, whereas that of a RAID-51 system is always less than 1/2. Consequently, these two systems cannot be fairly compared.

The MTTDL of a system comprising n_G RAID arrays is assessed by [8]

$$\text{MTTDL}_{\text{sys}} = \frac{\text{MTTDL}_{\text{RAID}}}{n_G} , \tag{53}$$

where $\text{MTTDL}_{\text{RAID}}$ denotes the MTTDL of a single RAID array.

A. RAID-5 vs. RAID-6

Let N_5 and N_6 be the sizes of a RAID-5 and a RAID-6 array, respectively. Assuming the same storage efficiency, we deduce from (6) and (7) that $N_6 = 2\,N_5$. Also, the user data stored in a RAID-6 array can also be stored in a system of $n_G = 2$ RAID-5 arrays. Using (9) and (53), the approximate MTTDL of the RAID-5 system, $\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}}$, is obtained as follows:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}} = \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}}}{2} \tag{54}$$

$$\approx \frac{\mu}{2N_5(N_5-1)\,\lambda^2} . \tag{55}$$

Also, the approximate MTTDL of the RAID-6 system, $\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}$, is obtained from (14) by setting $N = N_6 = 2N_5$:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}} = \text{MTTDL}^{\text{(approx)}}_{\text{RAID-6}} \tag{56}$$

$$\approx \frac{\mu^2}{2N_5(2N_5-1)(2N_5-2)\,\lambda^3} . \tag{57}$$

Using (55) and (57) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}}}{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}} = 2\,(2N_5-1)\cdot\frac{\lambda}{\mu} . \tag{58}$$

Thus, the reliability of the RAID-5 system is less than that of the RAID-6 system by a magnitude dictated by the ratio λ/μ, which is very small.

B. RAID-6 vs. RAID-51

Assuming the same storage efficiency for the two systems, we deduce from (7) and (22) that $N_6 = (4 - N_6)\,D$, which is satisfied by $N_6 = D = 3$ only. Furthermore, the user data stored in a RAID-51 array comprised of three pairs can also be stored in system of $n_G = 2$ RAID-6 arrays of size three. The approximate MTTDL of the RAID-6 array, $\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}$, is obtained from (14) by setting $N = 3$:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}} \approx \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-6}}}{2} \approx \frac{\mu^2}{12\,\lambda^3} . \tag{59}$$

The approximate MTTDL of the RAID-51 system, $\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-51}}$, is obtained from (34) by setting $D = 3$:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-51}} \approx \frac{\mu^3}{18\,\lambda^4} . \tag{60}$$

Using (59) and (60) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}}{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-51}}} = \frac{3}{2}\cdot\frac{\lambda}{\mu} . \tag{61}$$

Thus, the reliability of the RAID-6 system is lower than that of the RAID-51 system by a magnitude dictated by the ratio λ/μ, which is very small.

C. RAID-6 vs. 2D-RAID-5

In general, there are several combinations of N_6, K and D that yield the same storage efficiency for the two systems. From (7) and (37), it follows that

$$\frac{N_6-2}{N_6} = \frac{(K-1)\,(D-1)}{K\,D} , \tag{62}$$

which also implies that

$$D = \frac{(K-1)\,N_6}{2\,K - N_6} . \tag{63}$$

First, we examine whether there is a square 2D-RAID-5 system that has the same storage efficiency as that of a RAID-6 system. Substituting $D = K$ into (62), after some manipulations, yields $N_6 - K = K/(2K-1)$, which is not feasible given that $0 < K/(2K-1) < 1$, for $K > 1$. Therefore, we proceed by assuming, without loss of generality, that $D \geq K+1$. It follows that $K/(K+1) \leq D/(D-1) < 1$, which using (62) implies that $(K-1)/(K+1) \leq (N_6-2)/N_6 < (K-1)/K$, which in turn yields

$$K+1 \;\leq\; N_6 \;\leq\; 2\,K-1 . \tag{64}$$

Let us consider a system comprised of a single 2D-RAID-5 array with the user data stored in $(K-1)(D-1)$ devices. This data can also be stored in a system comprised of n_G RAID-6 arrays, where

$$n_G \;=\; \frac{(K-1)\,(D-1)}{N_6 - 2} \;\overset{(63)}{=}\; \frac{K\,(K-1)}{2\,K - N_6} \; . \qquad (65)$$

From (64) and (65), it follows that

$$K \;\le\; n_G \;\le\; K\,(K-1) \; . \qquad (66)$$

The values of K, D, and N_6 that that yield the same storage efficiency for the two systems are listed in Table II. We now fix K and consider the following two extreme combinations of the other two parameters, N_6 and D, obtained using (64) and (63), respectively.

1) $N_6 = D = K+1$*:* From (65), it follows that the user data is stored in a system of $n_G = K$ RAID-6 arrays. Using (14) and (53), the approximate MTTDL of the RAID-6 system, $\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}$, is obtained as follows:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}} \;=\; \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-6}}}{K} \qquad (67)$$

$$\approx\; \frac{\mu^2}{(K+1)\,K^2\,(K-1)\,\lambda^3} \; . \qquad (68)$$

TABLE II. Equal Storage Efficiency Values ($K \le 20$)

K	D	N_6	K	D	N_6	K	D	N_6
2	3	3	11	12	12	16	33	22
3	4	4	11	34	17	16	45	24
3	10	5	11	45	18	16	65	26
4	5	5	11	100	20	16	81	27
4	9	6	11	210	21	16	105	28
4	21	7	12	13	13	16	145	29
5	6	6	12	22	16	16	225	30
5	16	8	12	33	18	16	465	31
5	36	9	12	55	20	17	18	18
6	7	7	12	77	21	17	52	26
6	10	8	12	121	22	17	120	30
6	15	9	12	253	23	17	256	32
6	25	10	13	14	14	17	528	33
6	55	11	13	27	18	18	19	19
7	8	8	13	40	20	18	34	24
7	15	10	13	66	22	18	51	27
7	22	11	13	92	23	18	85	30
7	36	12	13	144	24	18	136	32
7	78	13	13	300	25	18	187	33
8	9	9	14	15	15	18	289	34
8	21	12	14	39	21	18	595	35
8	49	14	14	78	24	19	20	20
8	105	15	14	169	26	19	39	26
9	10	10	14	351	27	19	58	29
9	16	12	15	16	16	19	96	32
9	28	14	15	21	18	19	153	34
9	40	15	15	28	20	19	210	35
9	64	16	15	46	23	19	324	36
9	136	17	15	56	24	19	666	37
10	11	11	15	70	25	20	21	21
10	21	14	15	91	26	20	57	30
10	27	15	15	126	27	20	76	32
10	36	16	15	196	28	20	133	35
10	51	17	15	406	29	20	171	36
10	81	18	16	17	17	20	361	38
10	171	19	16	25	20	20	741	39

Also, the approximate MTTDL of the 2D-RAID-5 system, $\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}}$, is obtained from (52) as follows:

$$\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}} \;=\; \text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \qquad (69)$$

$$\approx\; \frac{2\,\mu^3}{3\,(K+1)\,K^2\,(K-1)\,\lambda^4} \; . \qquad (70)$$

Using (68) and (70) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}}{\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}}} \;=\; \frac{3}{2} \cdot \frac{\lambda}{\mu} \; . \qquad (71)$$

Thus, the reliability of the RAID-6 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the ratio λ/μ, which is very small.

2) $N_6 = 2K-1$ *and* $D = (K-1)(2K-1)$*:* From (65), it follows that the user data is stored in a system of $n_G = K(K-1)$ RAID-6 arrays. Using (14) and (53), the approximate MTTDL of the RAID-6 system is obtained as follows:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}} \;=\; \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-6}}}{K\,(K-1)} \qquad (72)$$

$$\approx\; \frac{\mu^2}{2\,K\,(K-1)^2\,(2K-1)\,(2K-3)\,\lambda^3} \; . \qquad (73)$$

Also, the approximate MTTDL of the 2D-RAID-5 system is obtained from (52) as follows:

$$\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}} \;=\; \text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \qquad (74)$$

$$\approx\; \frac{2\,\mu^3}{3\,K^2\,(K-1)^2\,(2K-1)\,(2K-3)\,\lambda^4} \; . \qquad (75)$$

Using (73) and (75) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-6}}}{\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}}} \;=\; \frac{3\,K}{4} \cdot \frac{\lambda}{\mu} \; . \qquad (76)$$

Thus, the reliability of the RAID-6 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the ratio λ/μ, which is very small.

D. RAID-5 vs. 2D-RAID-5

In general, there are several combinations of N_5, K and D that yield the same storage efficiency for the two systems. From (6) and (37), it follows that

$$\frac{N_5 - 1}{N_5} \;=\; \frac{(K-1)\,(D-1)}{K\,D} \; , \qquad (77)$$

or, equivalently,

$$\frac{2\,N_5 - 2}{2\,N_5} \;=\; \frac{(K-1)\,(D-1)}{K\,D} \; . \qquad (78)$$

From (62) and (78), it follows that the $(K,\, D,\, N_5)$ combinations correspond to the the $(K,\, D,\, N_6)$ ones, where N_6 is even and $N_5 = N_6/2$. From Table II, we deduce that the first two combinations are the following: $K=3$, $D=4$, $N_5=2$, and $K = 4$, $D = 9$, $N_5 = 3$. From (62), (64), and (78), it follows that

$$\frac{K+1}{2} \;\le\; N_5 \;\le\; K-1\,. \tag{79}$$

We now fix K to be an odd number and consider the following two extreme combinations regarding the other two parameters, D and N.

1) $N_5 = (K+1)/2$ and $D = K+1$: In this case, the user data is stored in $(K-1)(D-1) = K(K-1)$ devices in the 2D-RAID-5 system, which implies that this data can also be stored in a system of $n_G = 2K$ RAID-5 arrays. Using (9) and (53), the approximate MTTDL of the RAID-5 system is obtained as follows:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}} \;=\; \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}}}{2\,K} \tag{80}$$

$$\approx\; \frac{2\,\mu}{(K+1)\,K\,(K-1)\,\lambda^2}\,. \tag{81}$$

Also, the approximate MTTDL of the 2D-RAID-5 system is obtained from (52) as follows:

$$\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}} \;=\; \text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \tag{82}$$

$$\approx\; \frac{2\,\mu^3}{3\,(K+1)\,K^2\,(K-1)\,\lambda^4}\,. \tag{83}$$

Using (81) and (83) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}}}{\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}}} \;=\; 3\,K\,\cdot\,\left(\frac{\lambda}{\mu}\right)^2\,. \tag{84}$$

Thus, the reliability of the RAID-5 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the square of the ratio λ/μ, which is very small.

2) $N_5 = K-1$ and $D = (K-1)^2$: In this case, the user data is stored in $(K-1)(D-1) = (K-1)[(K-1)^2 - 1] = K(K-1)(K-2)$ devices in the 2D-RAID-5 system, which implies that this data can also be stored in a system of $n_G = K(K-1)$ RAID-5 arrays. Using (9) and (53), the approximate MTTDL of the RAID-5 system is obtained as follows:

$$\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}} \;=\; \frac{\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}}}{K\,(K-1)} \tag{85}$$

$$\approx\; \frac{\mu}{K\,(K-1)^2\,(K-2)\,\lambda^2}\,. \tag{86}$$

Also, the approximate MTTDL of the 2D-RAID-5 system is obtained from (52) as follows:

$$\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}} \;=\; \text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \tag{87}$$

$$\approx\; \frac{2\mu^3}{3\,K^2\,(K-1)^3\,(K-2)\,\lambda^4}\,. \tag{88}$$

Using (86) and (88) yields

$$\frac{\text{MTTDL}^{\text{(approx, system)}}_{\text{RAID-5}}}{\text{MTTDL}^{\text{(approx,system)}}_{\text{2D-RAID-5}}} \;=\; \frac{3\,K\,(K-1)}{2}\,\cdot\,\left(\frac{\lambda}{\mu}\right)^2\,. \tag{89}$$

Thus, the reliability of the RAID-5 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the square of the ratio λ/μ, which is very small.

E. Erasure Codes: Maximum Distance Separable (MDS) vs. non-MDS

An (l, m)-erasure code is a mapping from l user data symbols (or blocks) to a set of m ($> l$) symbols, called a codeword, in such a way that some subsets of the m blocks of the codeword can be used to decode the l user data blocks. Maximum distance separable (MDS) erasure codes have the property that *any* l of the m symbols can be used to decode a codeword. Examples of such codes include RAID-5 (an $(N, N+1)$-MDS code), RAID-6 (an $(N, N+2)$-MDS code), and r-way replication (an $(1, r)$-MDS code). As an MDS erasure code can decode data from any l of the m codeword symbols, a system employing such a code can sustain up to $(m-l)$ device failures. This implies that the most probable path to data loss has exactly $(m-l)$ 'hops', starting from the first-device failure and ending at data loss. As each hop has a probability proportional to λ/μ (when $\lambda/\mu \ll 1$), the resulting P_{DL} is proportional to $(\lambda/\mu)^{(m-l)}$. This can be seen from the P_{DL} equations (11) and (18) for RAID-5 and RAID-6, respectively.

In contrast, erasure codes that do not have the MDS property may not be able to sustain $(m-l)$ device failures. Examples of such non-MDS codes include RAID-51 (a $(D-1, 2D)$ non-MDS code) and 2D-RAID-5 (a $((D-1)\,(K-1), D\,K)$ non-MDS code). Both these non-MDS codes can sustain any three device failures; however, the fact that they have considerably higher redundancy may allow them to sustain certain other subsets of more than three devices, e.g., the failure of an entire row or column. Note that $(m-l)$ is equal to $D+1$ (≥ 3) and $D+K-1$ (≥ 3) for the RAID-51 and 2D-RAID-5, respectively, and therefore could be much higher than three for larger values of D and K. Despite this, these codes can sustain only up to three arbitrary device failures. This implies that the most probable path to data loss for RAID-51 and 2D-RAID-5 has exactly three hops, starting from the first-device failure and ending at data loss, and hence the resulting P_{DL} is proportional to $(\lambda/\mu)^3$. This can be seen from the P_{DL} equations (33) and (51) for RAID-51 and 2D-RAID-5, respectively.

Although it may seem that non-MDS codes are not useful because they provide a lower reliability than their MDS equivalents for the same storage efficiency, they have an advantage over MDS codes in the presence of correlated device failures that makes them valuable in practice. To see this, consider a system employing RAID-51, where each RAID-5 array is across D devices belonging to a different storage node. Such a system can sustain the failure of any node even though a node failure implies that all D devices belonging to that node are considered failed. Therefore, by carefully selecting the non-MDS code and the data placement, data can be protected from correlated failures.

VII. Most Probable Paths to Data Loss

In the preceding sections, we demonstrated that the reliability of systems comprised of highly reliable devices can be well approximated by considering the most likely paths that lead to data loss, namely, the shortest paths. These paths represent the smallest number of successive device failures that lead to data loss.

Here, we demonstrate that in general the shortest paths may not be the most likely paths that lead to data loss. We therefore extend our methodology to account for the most probable paths that lead to data loss. Clearly, the most probable paths are direct paths to data loss, i.e., paths without loops, but they may not be the shortest ones.

A. Unrecoverable or Latent Errors

When the storage devices are disks, in addition to disk failures, data loss may occur owing to errors in individual disk sectors that cannot be recovered with a reread or the sector-based error-correction code (ECC). Such media-related errors are referred to as unrecoverable or latent sector errors [6][8][11][18]. We proceed by considering the family of devices that exhibit such behavior. The occurrence of unrecoverable sector errors is particularly problematic when combined with device failures. For example, if a device fails in a RAID-5 array, the rebuild process must read all the data on the remaining devices to reconstruct the data lost. Consequently, an unrecoverable error on any of the operational devices would result in an irrecoverable loss of data. A similar problem occurs when two devices fail in a RAID-6 scheme. In this case, any unrecoverable sector errors encountered on the good devices during the rebuild process also lead to data loss.

The system reliability depends on the probability P_s of an unrecoverable error on a typical sector [6], as well as the sector size, S, and the device capacity, C_d. It in fact depends on the number of sectors in a device, n_s, which is given by

$$n_s \;=\; \frac{C_d}{S} \; . \tag{90}$$

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

B. A RAID-6 Array Under Latent Errors

The effect of unrecoverable sector errors in a RAID-6 system was analyzed in [6]. We proceed by briefly reviewing the Markov model developed to characterize the system behavior and capture the corresponding state transitions. The corresponding CTMC model shown in Fig. 7 is obtained from [6, Fig. 7] by setting $\mu_1 = \mu_2 = \mu$. The numbered states of the Markov model represent the number of failed devices. The DF and UF states represent a data loss due to a device failure and an unrecoverable sector failure, respectively.

When the first device fails, the array enters degraded mode, which corresponds to the transition from state 0 to state 1, with a transition rate of $N\lambda$. The rebuild of a sector of the failed device is performed based on up to $N-1$ corresponding sectors residing on the remaining devices. The rebuild fails when two or more of these sectors are in error. Consequently, the probability P_{recf} that a given sector of the failed device cannot be reconstructed is equal to the probability that two or more of the corresponding sectors residing in the remaining devices are in error and is given by [9, Eq. (12)]:

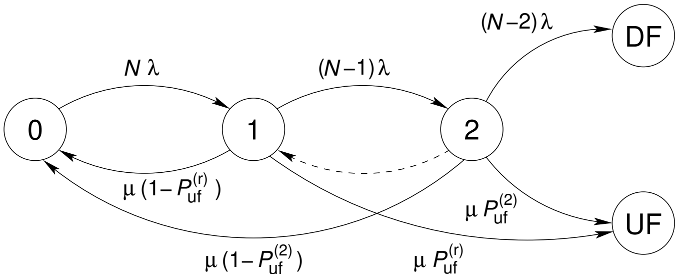

Figure 7. Reliability model for a RAID-6 array under latent errors.

$$P_{\text{recf}} \;\approx\; \binom{N-1}{2}\, P_s^2 \; . \tag{91}$$

Remark 3: Equation (91) is obtained from

$$P_{\text{recf}} \;=\; \sum_{j=2}^{N-1} \binom{N-1}{j} P_s^j (1-P_s)^{N-1-j} \;\approx\; \binom{N-1}{2} P_s^2 \; , \tag{92}$$

which is derived from Equation (47) of [6] by setting $P_{\text{seg}} = P_s$. Note that (92) accounts for all combinations of sector errors that cause the rebuild of the given sector to fail. However, the **most probable combinations of sector errors** are those that involve only two sectors in error, which is **the least number of sectors in error** that cause the rebuild to fail. These combinations yield the approximation given in (91).

The probability that an unrecoverable failure occurs in degraded mode because the rebuild of the failed device cannot be completed, $P_{\text{uf}}^{(\text{r})}$, is then given by [9, Eq. (9)]:

$$P_{\text{uf}}^{(\text{r})} = 1 - \left[1 - \binom{N-1}{2} P_s^2\right]^{n_s} \approx\; n_s \binom{N-1}{2} P_s^2 \, , \tag{93}$$

where n_s is the number of sectors in a device.

The system exits from state 1 owing to either another device failure or completion of the rebuild. The former event is represented by the state transition from state 1 to state 2 with a rate of $(N-1)\lambda$. The latter event occurs with a rate of μ and includes two possibilities: a failed rebuild (due to an unrecoverable failure) with probability $P_{\text{uf}}^{(\text{r})}$ and a successful rebuild with probability $1 - P_{\text{uf}}^{(\text{r})}$. The former event is represented by the state transition from state 1 to state UF with a rate of $\mu P_{\text{uf}}^{(\text{r})}$, and the latter one is represented by the state transition from state 1 to state 0 with a rate of $\mu(1 - P_{\text{uf}}^{(\text{r})})$.

When a second device fails (state transition from state 1 to state 2), the RAID-6 array enters the critical mode as an additional device failure leads to data loss. The rebuild of the two failed devices is performed based on the remaining $N-2$ devices. The rebuild fails if any sector of these $N-2$ devices is in error. The probability of this event, $P_{\text{uf}}^{(2)}$, is given by [9,

Eq. (10)]:

$$P_{\rm uf}^{(2)} = 1 - (1 - P_s)^{(N-2)\, n_s} \; \approx \; (N-2)\, n_s \, P_s \; . \qquad (94)$$

The system exits from state 2 owing to either another device failure or completion of the rebuild. The former event is represented by the state transition from state 2 to state DF with a rate of $(N-2)\lambda$. The latter event occurs with a rate of μ and includes two possibilities: a failed rebuild (due to an unrecoverable failure) with probability $P_{\rm uf}^{(2)}$ and a successful rebuild with probability $1 - P_{\rm uf}^{(2)}$. The former event is represented by the state transition from state 2 to state UF with a rate of $\mu P_{\rm uf}^{(2)}$, and the latter one by the state transition from state 2 to state 0 with a rate of $\mu(1 - P_{\rm uf}^{(2)})$.

We now proceed to show how the approximate MTTDL of the system can be derived in a straightforward manner by appropriate application of the direct-path-approximation technique. The transition from state 0 to state 1 represents the first device failure. The shortest path to data loss involves a subsequent transition from state 1 to UF, with a corresponding probability $P_{1\to{\rm UF}}$ given by

$$P_{1\to{\rm UF}} = \frac{\mu \, P_{\rm uf}^{(\rm r)}}{\mu + (N-1)\,\lambda} \overset{(4)}{\approx} P_{\rm uf}^{(\rm r)} \; . \qquad (95)$$

Note that there are two additional non-shortest paths, namely $1 \to 2 \to$ DF and $1 \to 2 \to$ UF, each involving two transitions, that lead to data loss. The probability $P_{1\to 2}$ of the transition from state 1 to state 2 is given by (15)

$$P_{1\to 2} = \frac{(N-1)\,\lambda}{\mu + (N-1)\,\lambda} \approx \frac{(N-1)\,\lambda}{\mu} \; . \qquad (96)$$

The probability $P_{2\to{\rm DF}}$ of the transition from state 2 to state DF is given by (16)

$$P_{2\to{\rm DF}} = \frac{(N-2)\,\lambda}{\mu + (N-2)\,\lambda} \approx \frac{(N-2)\,\lambda}{\mu} \; . \qquad (97)$$

Also, the probability $P_{2\to{\rm UF}}$ of the transition from state 2 to state UF is given by

$$P_{2\to{\rm UF}} = \frac{\mu \, P_{\rm uf}^{(2)}}{\mu + (N-2)\,\lambda} \approx P_{\rm uf}^{(2)} \; . \qquad (98)$$

Consequently, the probabilities of the two paths to data loss, $P_{1\to 2\to{\rm DF}}$ and $P_{1\to 2\to{\rm UF}}$, are given by

$$P_{1\to 2\to{\rm DF}} = P_{1\to 2}\, P_{2\to{\rm DF}} \approx \frac{(N-1)(N-2)\,\lambda^2}{\mu^2} \; , \qquad (99)$$

and

$$P_{1\to 2\to{\rm UF}} = P_{1\to 2}\, P_{2\to{\rm UF}} \approx \frac{(N-1)\,\lambda \, P_{\rm uf}^{(2)}}{\mu} \; . \qquad (100)$$

Thus, the probability of the direct paths to data loss that cross state 2, $P_{1\to 2\to{\rm DL}}$, is given by

$$\begin{aligned} P_{1\to 2\to{\rm DL}} &= P_{1\to 2\to{\rm DF}} + P_{1\to 2\to{\rm UF}} \\ &\approx \frac{(N-1)\,\lambda}{\mu} \left[\frac{(N-2)\,\lambda}{\mu} + P_{\rm uf}^{(2)} \right] \; . \end{aligned} \qquad (101)$$

From (95) and (99), and using (93), it follows that the ratio of the probabilities of the two paths $1 \to$ UF and $1 \to 2 \to$ DF is given by

$$\frac{P_{1\to{\rm UF}}}{P_{1\to 2\to{\rm DF}}} \approx \frac{1}{2}\, n_s \left(\frac{\lambda}{\mu}\right)^{-2} P_s^2 \; . \qquad (102)$$

Clearly, for very small values of P_s, this ratio is also very small, which implies that the path $1 \to 2 \to$ DF is significantly more probable than the shortest path $1 \to$ UF. In contrast to the cases considered in Sections IV and V, where the shortest paths were also the most probable ones, here the shortest path is not. **Therefore, we need to enhance the notion of the shortest paths by considering the most probable ones.**

In view of this finding, we proceed by assessing the system reliability in a region of small values of P_s in which the path $1 \to 2 \to$ DF is the most probable one. From (99) and (100), and using (94), it follows that the path $1 \to 2 \to$ DF is the most probable one when P_s is in region A obtained by

$$\frac{\lambda}{\mu} \gg n_s \, P_s \quad \Leftrightarrow \quad P_s \ll \frac{1}{n_s} \cdot \frac{\lambda}{\mu} \quad \text{(region A)} \; . \qquad (103)$$

Subsequently, for $P_s > \lambda/(n_s\, \mu)$, that is, when P_s is in region B, the other path to data loss, $1 \to 2 \to$ UF, becomes the most probable one. In fact, when $P_{\rm uf}^{(2)}$ approaches one, the $P_{\rm DL}$ and MTTDL no longer depend on P_s. Owing to (94), this occurs when P_s is in region C obtained by

$$P_{\rm uf}^{(2)} \approx 1 \quad \Leftrightarrow \quad P_s \gtrapprox \frac{1}{(N-2)\, n_s} \quad \text{(region C)} \; . \qquad (104)$$

Note that in region C, the path $1 \to 2 \to$ UF is also more probable than the shortest path $1 \to$ UF. Consequently, from (95) and (100), setting $P_{\rm uf}^{(2)} = 1$, and using (93), it follows that in region C it holds that

$$\begin{aligned} & n_s \binom{N-1}{2} P_s^2 \lessapprox (N-1)\, \frac{\lambda}{\mu} \\ & \qquad \Leftrightarrow \quad P_s \lessapprox \sqrt{\frac{2\,\lambda}{(N-2)\, n_s \, \mu}} \quad \text{(region C)} \; . \end{aligned} \qquad (105)$$

Subsequently, for $P_s \gtrapprox \sqrt{2\,\lambda/[(N-2)\, n_s\, \mu]}$, that is, when P_s is in region D, the shortest path to data loss $1 \to$ UF, becomes the most probable one. In fact, when $P_{\rm uf}^{(\rm r)}$ approaches one, $P_{\rm DL}$ and MTTDL no longer depend on P_s. Owing to (93), this occurs when P_s is in region E obtained by

$$P_{\rm uf}^{(\rm r)} \approx 1 \; \Leftrightarrow \; P_s \gtrapprox \sqrt{\frac{2}{(N-1)\,(N-2)\, n_s}} \quad \text{(region E)} \; . \qquad (106)$$

Combining the preceding, (101) yields

$$P_{\text{DL}} \approx \begin{cases} \frac{(N-1)(N-2)\,\lambda^2}{\mu^2} \,, & \text{A}: \; P_s \ll \frac{\lambda}{n_s\,\mu} \\ \frac{(N-1)(N-2)\,\lambda\,n_s}{\mu} \, P_s \,, & \text{B}: \; \frac{\lambda}{n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{(N-2)\,n_s} \\ \frac{(N-1)\,\lambda}{\mu} \,, & \text{C}: \; \frac{1}{(N-2)\,n_s} \lessapprox P_s \lessapprox \sqrt{\frac{2\,\lambda}{(N-2)\,n_s\,\mu}} \\ \frac{(N-1)(N-2)\,n_s}{2} \, P_s^2 \,, & \text{D}: \; \sqrt{\frac{2\,\lambda}{(N-2)\,n_s\,\mu}} \lessapprox P_s \lessapprox \sqrt{\frac{2}{(N-1)(N-2)\,n_s}} \\ 1 \,, & \text{E}: \; \sqrt{\frac{2}{(N-1)(N-2)\,n_s}} \lessapprox P_s \le 1 \,. \end{cases} \tag{107}$$

Substituting (107) into (3) yields

$$\text{MTTDL}_{\text{RAID-6}}^{(\text{approx})} \approx \begin{cases} \frac{\mu^2}{N(N-1)(N-2)\,\lambda^3} \,, & \text{A}: \; P_s \ll \frac{\lambda}{n_s\,\mu} \\ \frac{\mu}{N(N-1)(N-2)\,\lambda^2\,n_s} \, P_s^{-1} \,, & \text{B}: \; \frac{\lambda}{n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{(N-2)\,n_s} \\ \frac{\mu}{N(N-1)\,\lambda^2} \,, & \text{C}: \; \frac{1}{(N-2)\,n_s} \lessapprox P_s \lessapprox \sqrt{\frac{2\,\lambda}{(N-2)\,n_s\,\mu}} \\ \frac{2}{N(N-1)(N-2)\,\lambda\,n_s} \, P_s^{-2} \,, & \text{D}: \; \sqrt{\frac{2\,\lambda}{(N-2)\,n_s\,\mu}} \lessapprox P_s \lessapprox \sqrt{\frac{2}{(N-1)(N-2)\,n_s}} \\ \frac{1}{N\,\lambda} \,, & \text{E}: \; \sqrt{\frac{2}{(N-1)(N-2)\,n_s}} \lessapprox P_s \le 1 \,. \end{cases} \tag{108}$$

Remark 4: The preceding expression specifies three regions, namely A, D, and E, where the MTTDL is independent of P_s. This corresponds to three plateaus, as shown in [6, Fig. 9(c)].

Remark 5: Depending on the parameter values, some of the regions may vanish. For instance, region C vanishes when $\sqrt{2\,\lambda/[(N-2)\,n_s\,\mu]} < 1/[(N-2)n_s]$, or equivalently, $2\,(N-2)\,n_s\,\lambda/\mu < 1$.

Remark 6: The most probable paths are obtained by first identifying all direct paths to data loss, i.e., paths to data loss without loops, then evaluating their probability of occurrence, and finally selecting the most probable ones. Nevertheless, the MTTDL can be obtained analytically by considering all direct paths, which are more probable than those having loops. Therefore, it suffices to simply sum the probabilities of all direct paths to data loss to obtain the P_{DL}, and in turn, the MTTDL. The paths with the highest probabilities naturally dominate the sum and therefore implicitly determine the system reliability.

The direct paths to data loss are the following: 1 → UF, 1 → 2 → DF and 1 → 2 → UF. From (95), (99), and (100), it follows that

$$\begin{aligned} P_{\text{DL}} &\approx P_{1\to\text{UF}} + P_{1\to 2\to\text{DF}} + P_{1\to 2\to\text{UF}} \\ &\approx \min\left(1,\; P_{\text{uf}}^{(\text{r})} + \frac{(N-1)\,\lambda}{\mu}\left[\frac{(N-2)\,\lambda}{\mu} + P_{\text{uf}}^{(2)}\right]\right) . \end{aligned} \tag{109}$$

Note that the expression in (109) may exceed one and, as it expresses the probability of data loss, needs to be truncated to one. Assuming that the expression does not exceed one, substituting (109) into (3) yields

$$\text{MTTDL}_{\text{RAID-6}}^{(\text{approx})} \approx \frac{\mu^2}{N\{\mu^2\,P_{\text{uf}}^{(\text{r})} + (N-1)\,\lambda\,[(N-2)\,\lambda + \mu\,P_{\text{uf}}^{(2)}]\}\,\lambda} \,, \tag{110}$$

with $P_{\text{uf}}^{(\text{r})}$ and $P_{\text{uf}}^{(2)}$ given by (93) and (94), respectively.

We verify that by setting $\mu_1 = \mu_2 = \mu$, and using (4), the exact MTTDL expression given in [6, Eq. (52)] yields MTTDL $\approx \tau_0 \approx \mu^2/(N\lambda V)$, which after some manipulations gives the same result as in (110).

Remark 7: If the transition from state 2 to state 0 were not to state 0 but to state 1 instead, as shown in Fig. 7 by the dashed arrow, the corresponding MTTDL could still be approximated by (108) and (110) because the expressions for $P_{1\to\text{UF}}$, $P_{1\to 2\to\text{DF}}$, $p_{1\to 2\to\text{UF}}$, and P_{DL} given by (95), (99), (100), and (109) respectively, would still hold.

C. A Two-Dimensional RAID-5 Array Under Latent Errors

We consider the two-dimensional RAID-5 array analyzed in Section V-B in which the devices may contain unrecoverable or latent sector errors. We consider the probability P_s of an unrecoverable sector error to be small. This in turn implies that when considering the cases that lead to unsuccessful rebuilds of sectors residing on failed devices, and according to Remark 3, it suffices to consider only the most probable ones, which are those that involve the least number of sectors in error.

We now proceed to evaluate the MTTDL using the most-probable-path approximation. The transition from state $(0,0,0,0)$ to state $A \equiv (1,0,1,0)$ represents the first device failure as shown in in Fig. 8. The rebuild of the failed device can be performed using either the corresponding horizontal RAID-5 array H_1 or the corresponding vertical RAID-5 array V_1.

The rebuild of a given sector, say SEC, of the failed device fails when there are at least three corresponding sectors on other devices in error, with the four sectors (including SEC) occurring in a constellation of two horizontal rows (horizontal RAID-5 stripes) and two vertical columns (vertical RAID-5 stripes). Note that the sector in the constellation that resides opposite to SEC can be located in any of the $(K-1)(D-1)$ devices that are not in H_1 and V_1. Consequently, the probability P_{A} that SEC cannot be reconstructed is given by

$$P_{\text{sf}} \approx 1-(1-P_s^3)^{(K-1)(D-1)} \approx (K-1)(D-1)\,P_s^3 \,. \tag{111}$$

As the failed device contains n_s sectors, the probability that an unrecoverable failure occurs because its rebuild cannot be completed, P_{A}, is then given by

$$P_{\text{A}} \approx 1-(1-P_{\text{sf}})^{n_s} \overset{(111)}{\approx} 1-(1-P_s^3)^{(K-1)(D-1)n_s} \tag{112}$$

$$\approx (K-1)(D-1)\,n_s\,P_s^3 \,. \tag{113}$$

When a second device fails, the 2D-RAID-5 array enters either state $B \equiv (2,0,0,1)$, state $C \equiv (2,0,2,0)$ or state $D \equiv (0,1,2,0)$, as shown in Fig. 8. When the system is in state

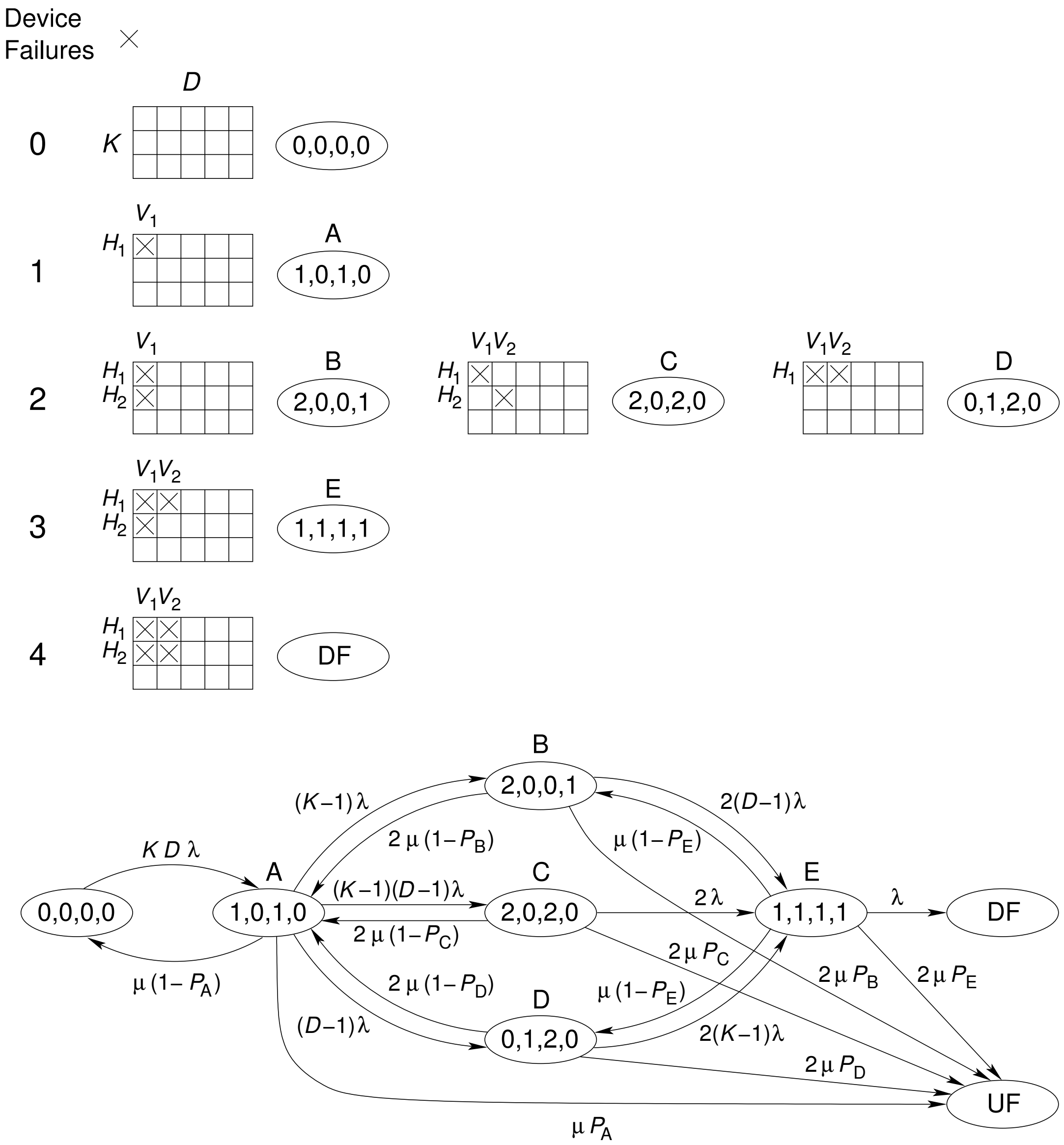

Figure 8. Reliability model for a 2D-RAID-5 array under latent errors.

B, the contents of the two failed devices in V_1 are rebuilt in parallel through the corresponding H_1 and H_2 RAID-5 arrays. The rebuild fails when there is a pair of corresponding sectors in error in some of the $D-1$ pairs of devices in the H_1 and H_2 RAID-5 arrays. As the number of such pairs is equal to $(D-1)\, n_s$, the probability of an unsuccessful rebuild, P_{B}, is given by

$$P_{\text{B}} \approx 1-(1-P_s^2)^{(D-1)n_s} \;\approx\; (D-1)\, n_s\, P_s^2 \; . \tag{114}$$

When the system is in state C, the contents of the two failed devices are rebuilt in parallel through the corresponding horizontal or vertical RAID-5 arrays. The rebuild fails when there is a pair of corresponding sectors in error in the (H_1, V_2) and (H_2, V_1) devices, where (H_1, V_2) refers to the device in row H_1 and column V_2, and (H_2, V_1) to that in row H_2 and column V_1. As the number of such pairs is equal to n_s, the probability of an unsuccessful rebuild, P_{C}, is given by

$$P_{\text{C}} \approx 1-(1-P_s^2)^{n_s} \;\approx\; n_s\, P_s^2 \; . \tag{115}$$

When the system is in state D, the contents of the two failed devices in H_1 are rebuilt in parallel through the corresponding V_1 and V_2 RAID-5 arrays. The rebuild fails when there is a pair of corresponding sectors in error in some of the $K-1$ pairs of devices in the V_1 and V_2 RAID-5 arrays. As the number of such pairs is equal to $(K-1)\, n_s$, the probability of an unsuccessful rebuild, P_{D}, is given by

$$P_{\text{D}} \approx 1-(1-P_s^2)^{(K-1)n_s} \;\approx\; (K-1)\, n_s\, P_s^2 \; . \tag{116}$$

When the system is in state $E \equiv (1,1,1,1)$, it suffices one latent sector error in the device in row H_2 and column V_2 to cause the rebuild to fail. As this device contains n_s sectors, the probability that an unrecoverable failure occurs because the rebuild of the failed devices cannot be completed, P_{E}, is then given by

$$P_{\text{E}} \approx 1-(1-P_s)^{n_s} \;\approx\; n_s\, P_s \; . \tag{117}$$

As shown in Fig. 8, the shortest path from state $(1,0,1,0)$

to data loss involves a subsequent transition from state $(1, 0, 1, 0)$ to UF, with a corresponding probability, $P_{\mathrm{A \to UF}}$, given by

$$P_{\mathrm{A \to UF}} = \frac{\mu \, P_{\mathrm{A}}}{\mu + (K\,D - 1)\,\lambda} \approx P_{\mathrm{A}} \; . \tag{118}$$

Next, we consider the three additional direct, two-hop non-shortest paths, namely, A → B → UF, A → C → UF, and A → D → UF, each involving two transitions, that lead to data loss. We proceed to evaluate the probabilities of their occurrence. From Fig. 8, it follows that

$$P_{\mathrm{B \to UF}} = \frac{2\,\mu \, P_{\mathrm{B}}}{2\,\mu + 2\,(D-1)\,\lambda} \approx P_{\mathrm{B}} \; , \tag{119}$$

$$P_{\mathrm{C \to UF}} = \frac{2\,\mu \, P_{\mathrm{C}}}{2\,\mu + 2\,\lambda} \approx P_{\mathrm{C}} \; , \tag{120}$$

and

$$P_{\mathrm{D \to UF}} = \frac{2\,\mu \, P_{\mathrm{D}}}{2\,\mu + 2\,(K-1)\,\lambda} \approx P_{\mathrm{D}} \; . \tag{121}$$

From (38) and (119), and using (114), it follows that

$$P_{\mathrm{A \to B \to UF}} = P_{\mathrm{A \to B}} \, P_{\mathrm{B \to UF}} \approx (K-1)\,\frac{\lambda}{\mu}\,P_{\mathrm{B}} \tag{122}$$

$$\approx (K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \; . \tag{123}$$

From (40) and (120), and using (115), it follows that

$$P_{\mathrm{A \to C \to UF}} = P_{\mathrm{A \to C}} \, P_{\mathrm{C \to UF}} \approx (K-1)\,(D-1)\,\frac{\lambda}{\mu}\,P_{\mathrm{C}} \tag{124}$$

$$\approx (K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \; . \tag{125}$$

From (42) and (121), and using (116), it follows that

$$P_{\mathrm{A \to D \to UF}} = P_{\mathrm{A \to D}} \, P_{\mathrm{D \to UF}} \approx (D-1)\,\frac{\lambda}{\mu}\,P_{\mathrm{D}} \tag{126}$$

$$\approx (K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \; . \tag{127}$$

Next, we consider the six additional direct non-shortest paths, namely, A → B → E → DF, A → C → E → DF, A → D → E → DF, A → B → E → UF, A → C → E → UF, and A → D → E → UF, each involving three transitions, that lead to data loss.

The probabilities of occurrence of the first three paths are the corresponding probabilities of these paths in the absence of sector errors given by (48), (49), and (50), respectively, that is,

$$\begin{aligned} P_{\mathrm{A \to B \to E \to DF}} &\approx P_{\mathrm{A \to C \to E \to DF}} \approx P_{\mathrm{A \to D \to E \to DF}} \\ &\approx \frac{(K-1)\,(D-1)}{2}\,\left(\frac{\lambda}{\mu}\right)^3 \; . \end{aligned} \tag{128}$$

Thus,

$$\begin{aligned} P_{\mathrm{A \to E \to DF}} &= P_{\mathrm{A \to B \to E \to DF}} + P_{\mathrm{A \to C \to E \to DF}} + P_{\mathrm{A \to D \to E \to DF}} \\ &\approx \frac{3\,(K-1)\,(D-1)}{2}\,\left(\frac{\lambda}{\mu}\right)^3 \; . \end{aligned} \tag{129}$$

According to (44), it holds that

$$P_{\mathrm{E \to DF}} \approx \frac{\lambda}{2\,\mu} \; . \tag{130}$$

From (129) and (130), we deduce that

$$P_{\mathrm{A \to E}} = 3\,(K-1)\,(D-1)\,\left(\frac{\lambda}{\mu}\right)^2 \; . \tag{131}$$

Also, it holds that

$$P_{\mathrm{E \to UF}} = \frac{2\,\mu \, P_{\mathrm{E}}}{2\,\mu + \lambda} \approx P_{\mathrm{E}} \; . \tag{132}$$

Combining (131) and (132) yields

$$P_{\mathrm{A \to E \to UF}} = P_{\mathrm{A \to E}} \, P_{\mathrm{E \to UF}} \approx 3\,(K-1)\,(D-1)\,\left(\frac{\lambda}{\mu}\right)^2 P_{\mathrm{E}} \; . \tag{133}$$

Thus, the probability of the direct paths to data loss that cross state E is given by

$$\begin{aligned} P_{\mathrm{A \to E \to DL}} &= P_{\mathrm{A \to E \to DF}} + P_{\mathrm{A \to E \to UF}} \\ &\approx 3\,(K-1)\,(D-1)\,\left(\frac{\lambda}{\mu}\right)^2 \left(\frac{\lambda}{2\,\mu} + P_{\mathrm{E}}\right) \; . \end{aligned} \tag{134}$$

From (118) and (129), and using (113), it follows that the ratio of the probabilities of the two paths A → UF and A → E → DF is given by

$$\frac{P_{\mathrm{A \to UF}}}{P_{\mathrm{A \to E \to DF}}} \approx \frac{2}{3}\,n_s\,\left(\frac{\lambda}{\mu}\right)^{-3} P_s^3 \; . \tag{135}$$

Clearly, for very small values of P_s, this ratio is also very small, which implies that the path A → E → DF is significantly more probable than the shortest path A → UF.

In view of this finding, we proceed by assessing the system reliability in a region of small values of P_s in which the path A → E → DF is the most probable one. Using (117), it follows that the first term of the summation in (134) dominates when P_s is in region H obtained by

$$\frac{\lambda}{2\mu} \gg n_s\,P_s \quad \Leftrightarrow \quad P_s \ll \frac{1}{2} \cdot \frac{1}{n_s} \cdot \frac{\lambda}{\mu} \quad \text{(region H)} \; . \tag{136}$$

Subsequently, for $P_s > \lambda/(2\,n_s\,\mu)$, that is, when P_s is in region I, the other path to data loss A → E → UF becomes the most probable one. In fact, when P_{E} approaches one, the P_{DL} and MTTDL no longer depend on P_s. Owing to (117), this occurs when P_s is in region J obtained by

$$P_{\mathrm{E}} \approx 1 \quad \Leftrightarrow \quad P_s \gtrapprox \frac{1}{n_s} \quad \text{(region J)} \; . \tag{137}$$

Note that in region J, the path A → E → UF is also more probable than any of the paths A → B → UF, A → C → UF, and A → D → UF. For small values of P_s, according to (123), (125), and (127), these two-hop paths are equally likely to occur. Therefore, the probability $P_{\mathrm{A \to X \to UF}}$ of a transition from state A to state UF through some other state X (X = B or C or D) is given by

$$P_{\mathrm{A \to X \to UF}} = 3\,(K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \; . \tag{138}$$

Consequently, from (133) and (138), and setting $P_{\text{E}} = 1$, it follows that in region J it holds that

$$3\,(K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \lessapprox 3\,(K-1)\,(D-1)\,\left(\frac{\lambda}{\mu}\right)^2$$
$$\Leftrightarrow \quad P_s \lessapprox \sqrt{\frac{\lambda}{n_s\,\mu}} \quad \text{(region J)} . \tag{139}$$

Subsequently, for $P_s \gtrapprox \sqrt{\lambda/(n_s\,\mu)}$, that is, when P_s is in region L, the two-hop paths to data loss, A $\rightarrow$ X $\rightarrow$ UF, become the most probable ones. In fact, as P_s increases, according to (114), (115), and (116), first P_{B}, then P_{D} and P_{C} approach one, and therefore the P_{DL} and MTTDL no longer depend on P_s. This occurs when P_s is in region M, with the corresponding probability, $P_{\text{A}\rightarrow\text{X}\rightarrow\text{UF}}$, obtained from (122), (124), and (126), by setting $P_{\text{B}} = P_{\text{C}} = P_{\text{D}} = 1$, that is,

$$P_{\text{A}\rightarrow\text{X}\rightarrow\text{UF}} = (K\,D-1)\,\frac{\lambda}{\mu} \quad \text{(region M)} . \tag{140}$$

Combining (138) and (140), we deduce that in region M it holds that

$$3\,(K-1)\,(D-1)\,n_s\,\frac{\lambda}{\mu}\,P_s^2 \gtrapprox (K\,D-1)\,\frac{\lambda}{\mu}$$
$$\Leftrightarrow \quad P_s \gtrapprox \sqrt{\frac{KD-1}{3\,(K-1)\,(D-1)\,n_s}} \quad \text{(region M)} . \tag{141}$$

Also, in region M, the paths A $\rightarrow$ X $\rightarrow$ UF are more probable than the shortest path A $\rightarrow$ UF. Consequently, from (118) and (140), and using (113), it follows that in region M it holds that

$$(K-1)\,(D-1)\,n_s\,P_s^3 \lessapprox (K\,D-1)\,\frac{\lambda}{\mu}$$
$$\Leftrightarrow \quad P_s \lessapprox \left[\frac{(K\,D-1)\,\lambda}{(K-1)\,(D-1)\,n_s\,\mu}\right]^{\frac{1}{3}} \quad \text{(region M)} . \tag{142}$$

Subsequently, for $P_s \gtrapprox \{(KD-1)\,\lambda/[(K-1)\,(D-1)\,n_s\,\mu]\}^{1/3}$, that is, when P_s is in region Q, the shortest path to data loss A $\rightarrow$ UF becomes the most probable one. In fact, when P_{A} approaches one, the P_{DL} and MTTDL no longer depend on P_s. Owing to (113), this occurs when P_s is in region R obtained by

$$P_{\text{A}} \approx 1 \;\Leftrightarrow\; P_s \gtrapprox \left[\frac{1}{(K-1)\,(D-1)\,n_s}\right]^{\frac{1}{3}} \quad \text{(region R)} . \tag{143}$$

Combining the preceding, (101) yields

$$P_{\text{DL}} \approx \begin{cases} \frac{3(K-1)(D-1)\,\lambda^3}{2\,\mu^3}\,, & \text{H}: \; P_s \ll \frac{\lambda}{2\,n_s\,\mu} \\ \frac{3(K-1)(D-1)\,\lambda^2\,n_s}{\mu^2}\,P_s\,, & \text{I}: \; \frac{\lambda}{2\,n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{n_s} \\ \frac{3(K-1)(D-1)\,\lambda^2}{\mu^2}\,, & \text{J}: \; \frac{1}{n_s} \lessapprox P_s \lessapprox \sqrt{\frac{\lambda}{n_s\,\mu}} \\ \frac{3(K-1)(D-1)\,\lambda\,n_s}{\mu}\,P_s^2\,, & \\ \qquad \text{L}: \; \sqrt{\frac{\lambda}{n_s\,\mu}} \lessapprox P_s \lessapprox \sqrt{\frac{KD-1}{3\,(K-1)\,(D-1)\,n_s}} & \\ \frac{(KD-1)\,\lambda}{\mu}\,, & \\ \qquad \text{M}: \; \sqrt{\frac{KD-1}{3\,(K-1)\,(D-1)\,n_s}} \lessapprox P_s \lessapprox \left[\frac{(KD-1)\,\lambda}{(K-1)(D-1)\,n_s\,\mu}\right]^{\frac{1}{3}} & \\ (K-1)(D-1)\,n_s\,P_s^3\,, & \\ \qquad \text{Q}: \; \left[\frac{(KD-1)\,\lambda}{(K-1)(D-1)\,n_s\,\mu}\right]^{\frac{1}{3}} \lessapprox P_s \lessapprox \left[\frac{1}{(K-1)(D-1)\,n_s}\right]^{\frac{1}{3}} & \\ 1\,, \; \text{R}: \; \left[\frac{1}{(K-1)(D-1)\,n_s}\right]^{\frac{1}{3}} \lessapprox P_s \leq 1\,. & \end{cases} \tag{144}$$

Substituting (144) into (3), and considering $N = KD$, yields the approximate MTTDL of the two-dimensional RAID-5 system, $\text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}}$, given by

$$\text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \approx \begin{cases} \frac{2\,\mu^3}{3K(K-1)D(D-1)\,\lambda^4}\,, & \text{H}: \; P_s \ll \frac{\lambda}{2\,n_s\,\mu} \\ \frac{\mu^2}{3K(K-1)D(D-1)\,\lambda^3\,n_s}\,P_s^{-1}\,, & \\ & \text{I}: \; \frac{\lambda}{2\,n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{n_s} \\ \frac{\mu^2}{3K(K-1)D(D-1)\,\lambda^3}\,, & \text{J}: \; \frac{1}{n_s} \lessapprox P_s \lessapprox \sqrt{\frac{\lambda}{n_s\,\mu}} \\ \frac{\mu}{3K(K-1)D(D-1)\,\lambda^2\,n_s}\,P_s^{-2}\,, & \\ \qquad \text{L}: \; \sqrt{\frac{\lambda}{n_s\,\mu}} \lessapprox P_s \lessapprox \sqrt{\frac{KD-1}{3\,(K-1)\,(D-1)\,n_s}} & \\ \frac{\mu}{KD(KD-1)\,\lambda^2}\,, & \\ \qquad \text{M}: \; \sqrt{\frac{KD-1}{3\,(K-1)\,(D-1)\,n_s}} \lessapprox P_s \lessapprox \left[\frac{(KD-1)\,\lambda}{(K-1)(D-1)\,n_s\,\mu}\right]^{\frac{1}{3}} & \\ \frac{1}{K(K-1)D(D-1)\,\lambda\,n_s}\,P_s^{-3}\,, & \\ \qquad \text{Q}: \; \left[\frac{(KD-1)\,\lambda}{(K-1)(D-1)\,n_s\,\mu}\right]^{\frac{1}{3}} \lessapprox P_s \lessapprox \left[\frac{1}{(K-1)(D-1)\,n_s}\right]^{\frac{1}{3}} & \\ \frac{1}{KD\lambda}\,, \qquad \text{R}: \; \left[\frac{1}{(K-1)(D-1)\,n_s}\right]^{\frac{1}{3}} \lessapprox P_s \leq 1\,. & \end{cases} \tag{145}$$

Following Remark 6, we obtain the P_{DL} by summing the probabilities of all direct paths to data loss. From (118), (122), (124), (126), and (134), it follows that

$$\begin{aligned} P_{\text{DL}} \approx \; & P_{\text{A}\rightarrow\text{UF}} + P_{\text{A}\rightarrow\text{B}\rightarrow\text{UF}} + P_{\text{A}\rightarrow\text{C}\rightarrow\text{UF}} + P_{\text{A}\rightarrow\text{D}\rightarrow\text{UF}} \\ & + P_{\text{A}\rightarrow\text{E}\rightarrow\text{DF}} + P_{\text{A}\rightarrow\text{E}\rightarrow\text{UF}} \; = \; \min\Bigg(1, \; P_{\text{A}} \\ & + [(K-1)P_{\text{B}} + (K-1)(D-1)\,P_{\text{C}} + (D-1)P_{\text{D}}]\,\frac{\lambda}{\mu} \\ & + 3\,(K-1)\,(D-1)\left(\frac{\lambda}{2\,\mu} + P_{\text{E}}\right)\left(\frac{\lambda}{\mu}\right)^2\Bigg) \; . \end{aligned} \tag{146}$$

Note that the expression in (146) may exceed one and, as it expresses the probability of data loss, needs to be truncated to one. Assuming that the expression does not exceed one, substituting (146) into (3), and considering $N = KD$, yields the approximate MTTDL of the two-dimensional RAID-5 array as follows:

$$\begin{aligned}&\text{MTTDL}^{\text{(approx)}}_{\text{2D-RAID-5}} \\ &\quad\approx \frac{1}{K\,D\,\lambda} \Bigg/ \Bigg\{ P_{\text{A}} + \\ &\quad + \left[(K-1)P_{\text{B}} + (K-1)(D-1)\,P_{\text{C}} + (D-1)P_{\text{D}}\right] \frac{\lambda}{\mu} \\ &\quad + 3\,(K-1)\,(D-1) \left(\frac{\lambda}{2\,\mu} + P_{\text{E}}\right) \left(\frac{\lambda}{\mu}\right)^2 \Bigg\} \ , \qquad (147)\end{aligned}$$

where P_{A}, P_{B}, P_{C}, P_{D}, and P_{E} are given by (112), (114), (115), (116), and (117), respectively.

D. A RAID-5 Array Under Latent Errors

The MTTDL of a RAID-5 array under latent sector errors was initially derived in [6] and is included in this article for completeness. The corresponding CTMC model is obtained from [6, Fig. 6] and shown in Fig. 9. When a device fails (state transition from state 0 to state 1), the RAID-5 array enters the critical mode as an additional device failure leads to data loss. The rebuild of the failed device is performed based on the remaining $N-1$ devices. The rebuild fails if any sector of these $N-1$ devices is in error. The probability of this event, $P^{(1)}_{\text{uf}}$, is given by [6, Eq. (1)]

$$P^{(1)}_{\text{uf}} = 1 - (1 - P_s)^{(N-1)\,n_s} \;\approx\; (N-1)\,n_s\,P_s \ . \qquad (148)$$

The system exits from state 1 owing to either another device failure or completion of the rebuild. The former event is represented by the state transition from state 1 to state UF with a rate of $\mu P^{(1)}_{\text{uf}}$, and the latter one by the state transition from state 1 to state 0 with a rate of $\mu(1 - P^{(1)}_{\text{uf}})$.

We now proceed to show how the approximate MTTDL of the system can be derived in a straightforward manner by appropriately applying the direct-path-approximation technique. The transition from state 0 to state 1 represents the first device failure. The probabilities of the direct paths to data loss, $P_{1\to\text{DF}}$ and $P_{1\to\text{UF}}$, are given by

$$P_{1\to\text{DF}} = \frac{(N-1)\,\lambda}{\mu + (N-1)\,\lambda} \approx \frac{(N-1)\,\lambda}{\mu} \ , \qquad (149)$$

and

$$P_{1\to\text{UF}} = \frac{\mu\,P^{(1)}_{\text{uf}}}{\mu + (N-1)\,\lambda} \approx P^{(1)}_{\text{uf}} \ . \qquad (150)$$

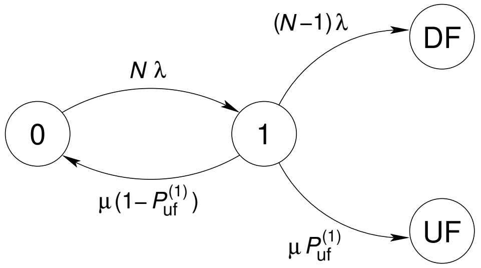

Figure 9. Reliability model for a RAID-5 array under latent errors.

Combining (149) and (150) yields

$$\begin{aligned}P_{\text{DL}} \approx P_{1\to\text{DL}} &= P_{1\to\text{DF}} + P_{1\to\text{UF}} \\ &= \frac{(N-1)\,\lambda + \mu\,P^{(1)}_{\text{uf}}}{\mu + (N-1)\,\lambda} \qquad (151) \\ &\approx \min\left(1,\ (N-1)\left(\frac{\lambda}{\mu}\right) + P^{(1)}_{\text{uf}}\right) \ . \qquad (152)\end{aligned}$$

Note that the expression in (152) may exceed one and, as it expresses the probability of data loss, needs to be truncated to one. Assuming that the expression does not exceed one, substituting (152) into (3) yields

$$\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}} \;\approx\; \frac{\mu}{N\,\lambda\,[(N-1)\,\lambda + \mu\,P^{(1)}_{\text{uf}}]} \ , \qquad (153)$$

which is the result obtained by Equation (45) of [6] when the first term of the nominator is ignored.

From (149) and (150), and using (148), it follows that the path 1 → DF is the most probable one when P_s is in region A obtained by

$$\frac{\lambda}{\mu} \gg n_s\,P_s \quad \Leftrightarrow \quad P_s \ll \frac{1}{n_s} \cdot \frac{\lambda}{\mu} \quad \text{(region A)} \ . \qquad (154)$$

Note that this is the same region as that in (103) for a RAID-6 array.

Subsequently, for $P_s > \lambda/(n_s\,\mu)$, that is, when P_s is in region F, the other path to data loss, 1 → UF, becomes the most probable one. In fact, when $P^{(1)}_{\text{uf}}$ approaches one, the P_{DL} and MTTDL no longer depend on P_s. Owing to (148), this occurs when P_s is in region G obtained by

$$P^{(2)}_{\text{uf}} \approx 1 \quad \Leftrightarrow \quad P_s \gtrapprox \frac{1}{(N-1)\,n_s} \quad \text{(region G)} \ . \qquad (155)$$

Combining the preceding, (152) yields

$$\begin{aligned}&P_{\text{DL}} \approx P_{1\to\text{DL}} \\ &\quad\approx \begin{cases} \frac{(N-1)\,\lambda}{\mu} \ , & \text{A}: \ P_s \ll \frac{\lambda}{n_s\,\mu} \\ (N-1)\,n_s\,P_s \ , & \text{F}: \ \frac{\lambda}{n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{(N-1)\,n_s} \\ 1 & \text{G}: \ \frac{1}{(N-1)\,n_s} \lessapprox P_s \le 1 \ . \end{cases} \qquad (156)\end{aligned}$$

Substituting (156) into (3) yields

$$\begin{aligned}&\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}} \approx \\ &\quad\approx \begin{cases} \frac{\mu}{N(N-1)\,\lambda^2} \ , & \text{A}: \ P_s \ll \frac{\lambda}{n_s\,\mu} \\ \frac{1}{N(N-1)\,\lambda\,n_s}\,P_s^{-1} \ , & \text{F}: \ \frac{\lambda}{n_s\,\mu} \lessapprox P_s \lessapprox \frac{1}{(N-1)\,n_s} \\ \frac{1}{N\,\lambda} & \text{G}: \ \frac{1}{(N-1)\,n_s} \lessapprox P_s \le 1 \ . \end{cases} \qquad (157)\end{aligned}$$

E. RAID-5 vs. RAID-6 Under Latent Errors

In region A, the ratio of the corresponding reliabilities is given by (58). In regions G and C, the corresponding MTTDLs are obtained from (157) and (108) as follows:

$$\text{MTTDL}^{\text{(approx)}}_{\text{RAID-5}} \;\approx\; \frac{1}{N_5\,\lambda} \quad \text{(region G)} \ , \qquad (158)$$

and

$$\text{MTTDL}^{\text{(approx)}}_{\text{RAID-6}} \;\approx\; \frac{\mu}{N_6\,(N_6-1)\,\lambda^2} \quad \text{(region C)} \ . \qquad (159)$$

Moreover, according to (54) and (56), it holds that $\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx)}}/2$ and $\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}}$, respectively. Consequently, from (158) and (159), and given that $N_6 = 2\,N_5$, it follows that in region C∩G it holds that

$$\frac{\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx, system)}}}{\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}}} = (N_6 - 1) \cdot \frac{\lambda}{\mu} \,. \tag{160}$$

Thus, the reliability of the RAID-5 system in region C∩G is less than that of the RAID-6 system by a magnitude dictated by the ratio λ/μ, which is very small. As this holds in both regions A and C∩G, we deduce that this also holds in region B∩F. Consequently, for all realistic values of P_s, the reliability of the RAID-5 system is lower than that of the RAID-6 system by a magnitude dictated by the ratio λ/μ.

F. RAID-6 vs. 2D-RAID-5 Under Latent Errors

In region H∩A, the ratios of the corresponding reliabilities are given by (71) and (76) for cases 1 and 2, respectively.

1) $N_6 = D = K+1$: In regions C and J, the corresponding MTTDLs are obtained from (108) and (145) as follows:

$$\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}} \approx \frac{\mu}{(K+1)\,K\,\lambda^2} \quad \text{(region C)}\,, \tag{161}$$

and

$$\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}} \approx \frac{\mu^2}{3\,(K+1)\,K^2\,(K-1)\,\lambda^3} \quad \text{(region J)}\,. \tag{162}$$

Also, according to (67) and (69), it holds that $\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}}/K$ and $\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}} = \mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}}$, respectively. Consequently, from (161) and (162), it follows that in region J∩C it holds that

$$\frac{\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}}}{\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}}} = 3\,(K-1) \cdot \frac{\lambda}{\mu} \,. \tag{163}$$

2) $N_6 = 2K-1$ and $D = (K-1)(2K-1)$: In regions C and J, the corresponding MTTDLs are obtained from (108) and (145) as follows:

$$\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}} \approx \frac{\mu}{2\,(K-1)\,(2K-1)\,\lambda^2} \quad \text{(region C)}\,, \tag{164}$$

and

$$\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}} \approx \frac{\mu^2}{3\,K^2\,(K-1)^2\,(2K-1)\,(2K-3)\,\lambda^3} \quad \text{(region J)}\,. \tag{165}$$

Also, according to (72) and (74), it holds that $\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}}/[K(K-1)]$ and $\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}} = \mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}}$, respectively. Consequently, from (164) and (165), it follows that in region J∩C it holds that

$$\frac{\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}}}{\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}}} = \frac{3\,K\,(2K-3)}{2} \cdot \frac{\lambda}{\mu} \,. \tag{166}$$

Thus, in both cases, the reliability of the RAID-6 system in region J∩C is lower than that of the 2D-RAID-5 system by a magnitude dictated by the ratio λ/μ, which is very small. As this holds in both regions H∩A and J∩C, we deduce that this also holds in region I∩B. Consequently, for all realistic values of P_s, the reliability of the RAID-6 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the ratio λ/μ.

G. RAID-5 vs. 2D-RAID-5 Under Latent Errors

In region H∩A, the ratios of the corresponding reliabilities are given by (84) and (89) for cases 1 and 2, respectively.

1) $N_5 = (K+1)/2$ and $D = K+1$: In region G, the corresponding MTTDL is obtained from (157) as follows:

$$\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx)}} \approx \frac{2}{(K+1)\,\lambda} \quad \text{(region G)}\,, \tag{167}$$

In region J, the corresponding MTTDL is given by (162). Also, according to (80) and (82), it holds that $\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx)}}/(2K)$ and $\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}} = \mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}}$, respectively. Consequently, from (167) and (162), it follows that in region J∩G it holds that

$$\frac{\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx, system)}}}{\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}}} = 3\,K\,(K-1) \cdot \left(\frac{\lambda}{\mu}\right)^2 \,. \tag{168}$$

2) $N_5 = K-1$ and $D = (K-1)^2$: In regions G and J, the corresponding MTTDLs are obtained from (157) and (145) as follows:

$$\mathrm{MTTDL}_{\text{RAID-5}}^{\text{(approx)}} \approx \frac{1}{(K-1)\,\lambda} \quad \text{(region G)}\,, \tag{169}$$

and

$$\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}} \approx \frac{\mu^2}{3\,K^2\,(K-1)^3\,(K-2)\,\lambda^3} \quad \text{(region J)}\,. \tag{170}$$

Also, according to (85) and (87), it holds that $\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}} = \mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx)}}/[K(K-1)]$ and $\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}} = \mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx)}}$, respectively. Consequently, from (169) and (170), it follows that in region J∩G it holds that

$$\frac{\mathrm{MTTDL}_{\text{RAID-6}}^{\text{(approx, system)}}}{\mathrm{MTTDL}_{\text{2D-RAID-5}}^{\text{(approx,system)}}} = 3\,K\,(K-1)\,(K-2) \cdot \left(\frac{\lambda}{\mu}\right)^2 \,. \tag{171}$$

Thus, in both cases, the reliability of the RAID-5 system in region J∩G is lower than that of the 2D-RAID-5 system by a magnitude dictated by the square of the ratio λ/μ, which is very small. As this holds in both regions H∩A and J∩G, we deduce that this also holds in region I∩F. Consequently, for all realistic values of P_s, the reliability of the RAID-5 system is lower than that of the 2D-RAID-5 system by a magnitude dictated by the square of the ratio λ/μ.

VIII. Numerical Results

We consider a system comprised of devices with $C_d = 1$ TB and $S = 512$ bytes. Note that in all cases P_{DL} depends on λ and μ only through their ratio λ/μ. Consequently, the quantity $\lambda\,\text{MTTDL}$, which owing to (3) and (53), is given by

$$\lambda\,\text{MTTDL} \approx \frac{1}{n_G\, N\, P_{\text{DL}}} \,. \qquad (172)$$

also depends on λ and μ only through their ratio λ/μ. In the remainder, we set $\lambda/\mu = 0.001$. The $\lambda\,$MTTDLs of RAID-6, 2D-RAID-5, and RAID-5 systems are evaluated analytically using (110), (147), and (153), respectively.

The combined effects of device and unrecoverable failures in a RAID-5 and a RAID-6 array ($n_G = 1$) of size $N = 8$ can be seen in Fig. 10 as a function of the unrecoverable sector error probability: it shows the most probable paths that lead to data loss along with the resulting $\lambda\,$MTTDL measure. The backward arrows have been included because they affect the probability of occurrence of these paths. The dotted backward arrows indicate transitions that are no longer possible.

The $\lambda\,$MTTDL for the RAID-5 array, indicated by the dashed line, exhibits two plateaus that, according to (157), correspond to the regions A and G. The first plateau, in region A, corresponds to the case where there are no unrecoverable errors and therefore data loss occurs owing to two successive device failures. The second plateau, in region G, corresponds to the first device failure after a mean time of $N\,\lambda$, which in turn leads to data loss during rebuild due to unrecoverable errors.

The $\lambda\,$MTTDL for the RAID-6 array, indicated by the solid line, exhibits three plateaus that, according to (108), correspond to the regions A, C, and G. The first plateau, in region A, corresponds to the case where there are no unrecoverable sector errors, and therefore data loss occurs owing to three successive device failures. In this case, the most probable path is not the shortest path, $1 \to$ UF, but the path $1 \to 2 \to$ DF, indicated by the solid red line in Fig. 11. This line is horizontal because, according to (99), the probability of occurrence of this path does not depend on P_s. In region B, the most probable path is the path $1 \to 2 \to$ UF, indicated by the dashed blue line in Fig. 11. Also, according to (100) and (104), in region C, the probability of occurrence of this path becomes independent of P_s, which results in the second plateau. This corresponds to a second device failure, which in turn leads to data loss during rebuild due to unrecoverable errors. Subsequently, in region D, the most probable path is the shortest path, $1 \to$ UF, indicated by the dotted green line in Fig. 11. Also, according to (95) and (106), in region E, the probability of occurrence of this path becomes one, independent of P_s, which results in the third plateau. This corresponds to the first device failure, which in turn leads to data loss during rebuild due to unrecoverable errors.

Note that the plateaus G and E correspond to the same MTTDL value of $1/(N\,\lambda)$. Similarly, the plateaus A and C correspond to the same MTTDL value of $\mu/[N(N-1)\,\lambda^2]$. From (103), (104), (154), and (155), it follows that region B is about the same as region F. Furthermore, from (108) and (157), it follows that in regions A, B, and C, the MTTDL of the RAID-5 array is lower than that of the RAID-6 array by a factor of $(N-2)\lambda/\mu$, $(N-2)\lambda/\mu$, and $(N-1)\lambda/\mu$, respectively.

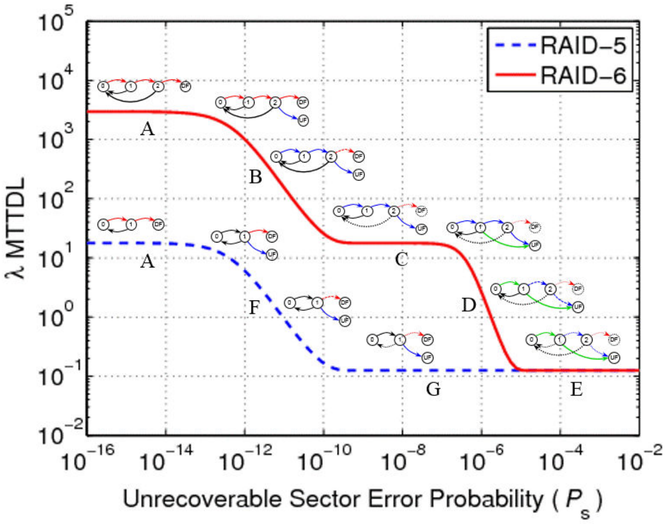

Figure 10. $\lambda\,$MTTDL for a RAID-5 and a RAID-6 array under latent errors ($\lambda/\mu = 0.001$, $N_5 = N_6 = 8$, and $C_d = 1$ TB).

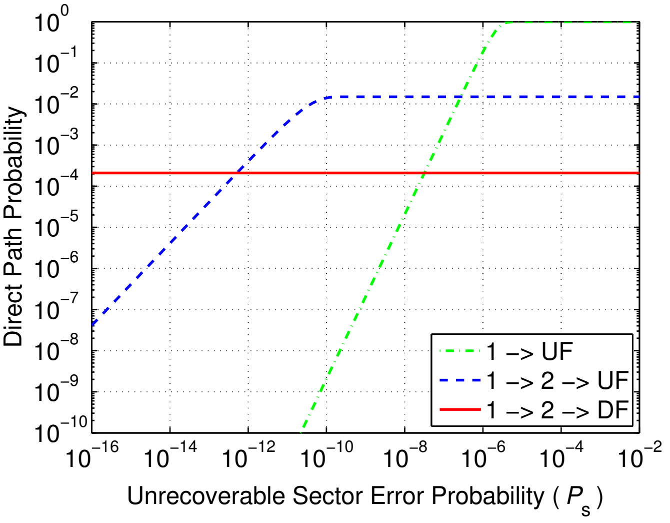

Figure 11. Direct-path probabilities for a RAID-6 array under latent errors ($\lambda/\mu = 0.001$, $N = 8$, and $C_d = 1$ TB).

Next, we consider a 2D-RAID-5 array with $K = 9$ and $D = 64$, and therefore a corresponding storage efficiency of 0.875. We also consider a system comprised of $n_G = 72$ RAID-5 arrays of size $N = 8$ and a system comprised of $n_G = 36$ RAID-6 arrays of size $N = 16$, such that these systems store the same amount of user data as the 2D-RAID-5 array under the same storage efficiency. The combined effects of device and unrecoverable failures on the $\lambda\,$MTTDL measure are shown in Fig. 12 as a function of the unrecoverable sector error probability. The various regions and plateaus are also depicted. The probabilities of occurrence of all direct paths to data loss for the 2D-RAID-5 array are shown in Fig. 13. We observe that the shortest path to data loss, A $\to$ UF, indicated by the dotted green line, becomes the most probable one only if

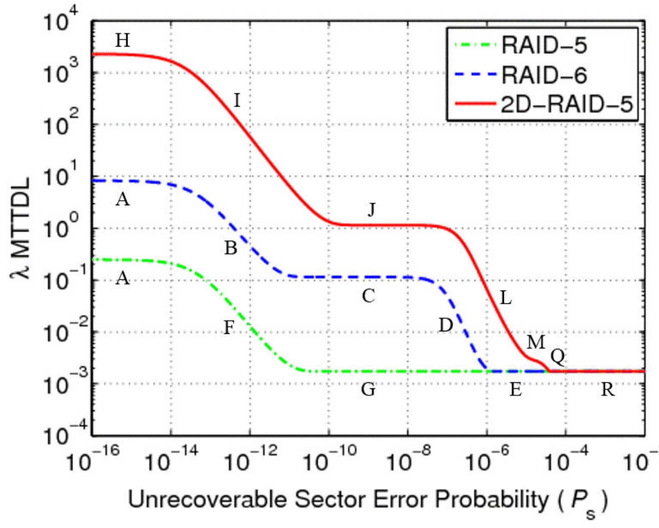

Figure 12. λ MTTDL for a RAID-5, RAID-6, and 2D-RAID-5 system under latent errors ($\lambda/\mu = 0.001$, $N_5 = 8$, $N_6 = 16$, $K = 9$, $D = 64$, and $C_d = 10$ TB).

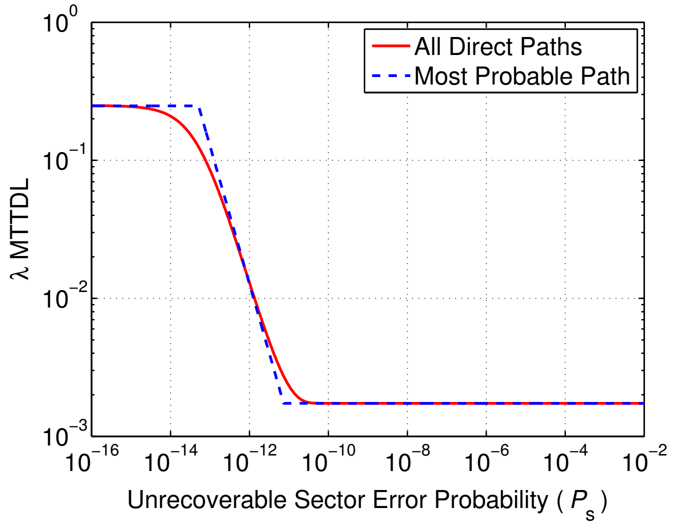

Figure 14. λ MTTDL for a RAID-5 system under latent errors ($\lambda/\mu = 0.001$, $N = 8$, $n_G = 72$, and $C_d = 10$ TB).

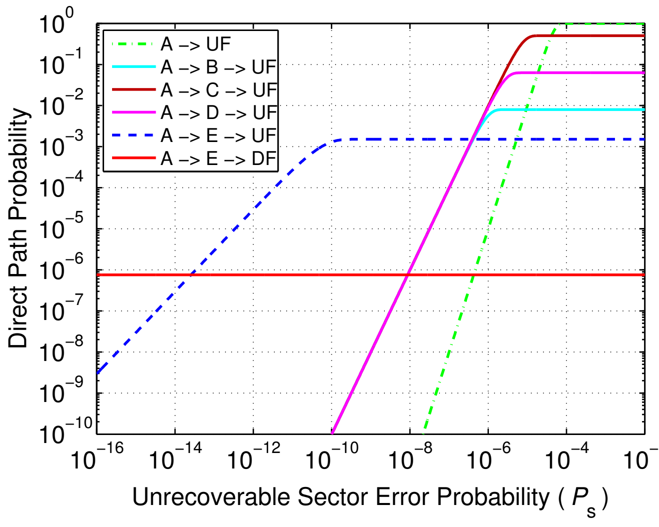

Figure 13. Direct-path probabilities for a 2D-RAID-5 array under latent errors ($\lambda/\mu = 0.001$, $K = 9$, $D = 64$, and $C_d = 10$ TB).

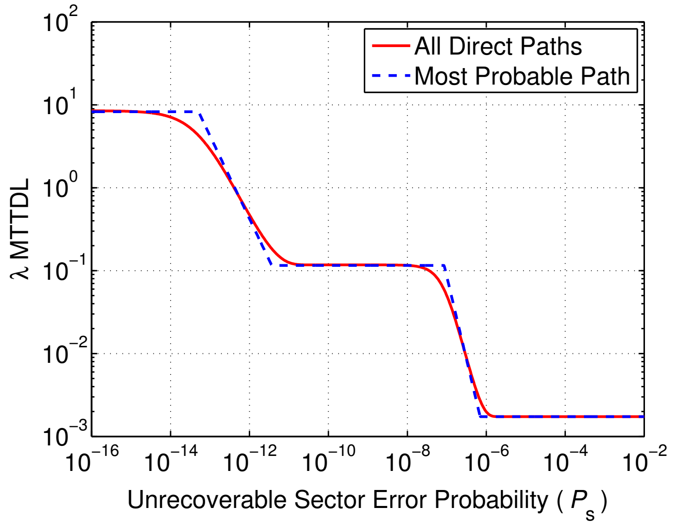

Figure 15. λ MTTDL for a RAID-6 system under latent errors ($\lambda/\mu = 0.001$, $N = 16$, $n_G = 36$, and $C_d = 10$ TB).

$P_s > 10^{-4}$. We also observe that for $P_s < 10^{-7}$, the reliability of the 2D-RAID-5 system is higher than that of the RAID-6 system, which in turn is higher than that of the RAID-5 system.

In Section VII, the MTTDL was derived in two ways: by considering the most probable path to data loss and by considering all direct paths to data loss (see Remark 6). The corresponding results for the RAID-5 system are obtained by (157) and (153) and shown in Fig. 14. The corresponding results for the RAID-6 system are obtained by (108) and (110) and shown in Fig. 15. Finally, the corresponding results for the 2D-RAID-5 array are obtained by (145) and (147) and shown in Fig. 16.

IX. Conclusions

We considered the mean time to data loss (MTTDL) metric, which assesses the reliability level of storage systems. This work presented a simple, yet efficient methodology to approximately assess it analytically for systems with highly reliable devices and a broad set of redundancy schemes. We extended the direct-path approximation to a more general method that considers the most probable paths, which are often the shortest paths, that lead to data loss. We subsequently applied this method to obtain a closed-form expression for the MTTDL of a RAID-51 system. We also considered a specific instance of a RAID-51 system, then derived the corresponding exact MTTDL, and subsequently confirmed that it matches that obtained from the shortest-path-approximation method. Closed-form approximations were also obtained for the MTTDL of

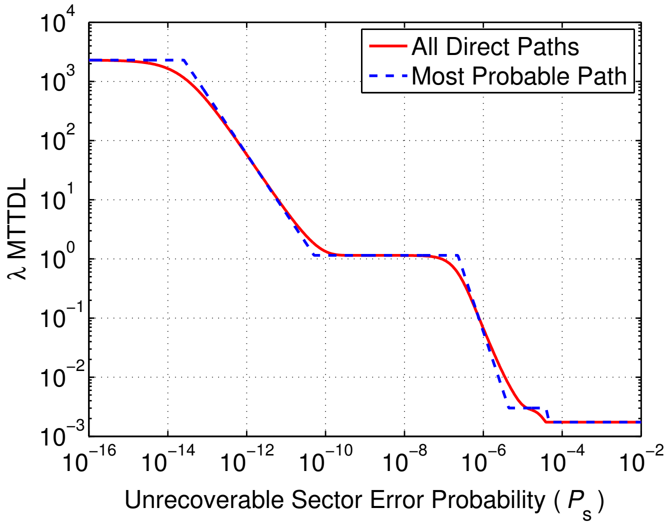

Figure 16. λMTTDL for a 2D-RAID-5 array under latent errors ($\lambda/\mu = 0.001$, $K = 9$, $D = 64$, and $C_d = 10$ TB).

RAID-6 and two-dimensional RAID-5 systems in the presence of unrecoverable errors and device failures. Subsequently, a thorough comparison of the reliability levels achieved by the redundancy schemes considered was conducted. As the direct-path approximation accurately predicts the reliability of non-Markovian systems with a single shortest path, we conjecture that the shortest-path-approximation method would also accurately predict the reliability of non-Markovian systems with multiple shortest paths.

Application of the shortest-path-approximation methodology developed to derive the MTTDL for systems using other redundancy schemes, such as erasure codes, is a subject of future work.

This methodology can also be applied to system models that additionally consider node, rack, and data-center failures. In such models, there may be short paths to data loss that are not very likely to occur (e.g., disaster events), and direct paths to data loss that are highly probable, but not necessarily short.

References

[1] 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) (Barcelona, Spain), Apr. 2015, pp. 6–12.

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

[5] 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.

[6] A. Dholakia, E. Eleftheriou, X.-Y. Hu, I. Iliadis, J. Menon, and K. Rao, "A new intra-disk redundancy scheme for high-reliability RAID storage systems in the presence of unrecoverable errors," ACM Trans. Storage, vol. 4, no. 1, May 2008, pp. 1–42.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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) (San Diego, CA), Apr. 2003, pp. 146–156.

[12] Q. Xin, T. J. E. Schwarz, and E. L. Miller, "Disk infant mortality in large storage systems," in Proceedings of the 13th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS) (Atlanta, GA), Sep. 2005, pp. 125–134.

[13] A. Wildani, T. J. E. Schwarz, E. L. Miller, and D. D. E. Long, "Protecting against rare event failures in archival systems," in Proceedings of the 17th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS) (London, UK), Sep. 2009, pp. 1–11.

[14] M. Bouissou and Y. Lefebvre, "A path-based algorithm to evaluate asymptotic unavailability for large markov models," in Proceedings of the 48th Annual Reliability and Maintainability Symposium, Jan. 2002, pp. 32–39.

[15] I. B. Gertsbakh, "Asymptotic methods in reliability theory: A review," Adv. App. Probability, vol. 16, no. 1, Mar. 1984, pp. 147–175.

[16] 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.

[17] 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.

[18] 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.

[19] ——, "Effect of latent errors on the reliability of data storage systems," in Proceedings of the 21th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Aug. 2013, pp. 293–297.

[20] 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) (Austin, TX), Dec. 2012, pp. 324–331.

[21] 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) (Paris, France), Sep. 2014, pp. 375–384.

[22] A. Thomasian, "Shortcut method for reliability comparisons in RAID," J. Syst. Software, vol. 79, no. 11, Nov. 2006, pp. 1599–1605.