Redundancy and Aging of Efficient Multidimensional MDS Parity-Protected Distributed Storage Systems

Suayb S. Arslan, *Member, IEEE*

***Abstract*—The effect of redundancy on the aging of an efficient maximum distance separable (MDS) parity-protected distributed storage system that consists of MDS-parity protected array of storage units is explored. In light of the experimental evidences and survey data, this paper develops generalized expressions for the reliability of array storage systems based on more realistic time to failure distributions such as Weibull. For instance, a distributed disk array system is considered in which the array components are disseminated across the network and are subject to independent failure rates. Based on such, generalized closed-form hazard rate expressions are derived. These expressions are extended to estimate the asymptotical reliability behavior of large-scale storage networks equipped with MDS parity-based protection. Unlike previous studies, a generic hazard rate function is assumed, a generic MDS code for parity generation is used, and an evaluation of the implications of adjustable redundancy level for an efficient distributed storage system is presented. Results of this paper are applicable to any erasure correction code as long as it is accompanied with a suitable structure and an appropriate encoding/decoding algorithm such that the MDS property is maintained.**

***Index Terms*—Aging, big data management, error correction coding, hazard rate, redundant array of inexpensive/independent disks (RAID), reliability.**

I. Introduction

ONE of the well known problems associated with parity-based redundant array of inexpensive disk (RAID) systems [1] is their vulnerability against multiple disk failures, mostly after which a restore mechanism is initiated and subsequent read errors inevitably occur due to lot of repeated reads. Similar trends can be observed in arrays of solid state drives (known as RAIS) for mass storage applications [2]. The statistical likelihood of multiple drive failures has never been a significant issue in the past. Over the years however, with the advanced technology, drives of few terabyte capacities are now put on sale. The scale of storage systems continues to grow to store peta-bytes of data and the likelihood of multiple drive failures become a reality. This led to the development of error checking and validation routines to maintain the data integrity. Conventional approach for data retention was to address the big data protection shortcomings of RAID by replication, a technique of making additional copies of data to avoid unrecoverable errors and lost data. Organizations also used replication schemes to help with failure scenarios, such as location specific failures, power outages, bandwidth unavailability, and so forth. However, as the size of the stored data scales up, the number of copies of the data required for robust protection grows. This increases the amount of inefficiency by adding extra cost to the overall system. Since replication leads to extremely inefficient use of system resources, parity-based protection using error correcting codes is more popular.

Drive failures can be regarded as arrivals of a renewal process characterized by a rate parameter. The drive failure rate, using a homogenous Poisson process, is the reciprocal of the mean time to failure (MTTF) numbers reported by the drive manufactures [3]. One of the earliest studies of the reliability analysis for disk array systems considered various RAID hierarchies and hot spots [4]. In a number of successive works, stripping is used to provide cost-effective I/O systems [1], [5] for seemless and reliable access to the user data. Most of the previous research modelings were based on single or double–parity schemes such as RAID 5 or RAID 6 in which maximum distance separable (MDS) codes are used for storage efficiency. MDS codes have the nice property that for a given array and parity size they allow maximum amount of recovery [6]. However, these studies mostly assume a Poisson process characterized by a constant failure rate of small size and cost–effective disk components. Unfortunately, these set of assumptions are shown to be unrealistic [7], [8]. In fact, an interesting observation is that the failure rates are rarely constant [8], [9].

There have been efforts in industry as well as in academia for accurately predicting the reliability of large-scale storage systems in terms of mean lifetime to failure rates. For example, an accurate yet complicated model is developed to include catastrophic failures and usage dependent data corruptions in [10]. The authors specifically pointed out that component failure rates have little, if not any, to share with the failure rate of the whole storage system. The times between successive system failures are reported to be relatively larger than what conventional models suggest, even though each component disk hazard rate is increasing [11]. Disk scrubbing is introduced and used in [12] as a remedy for latent defects that are usually independent of the size, use and the operation of disks. The latter study also uses homogenous Poisson model for reliability estimations.

It is clear that excessive failures (failures beyond the correction capability of the system) in any storage system are of particular interest because they may cause both unavailability and permanent data loss. On the other hand, the trend in the market is to grow the scale of distributed storage arrays in which the capacity as well as the reliability of each storage component almost double every year. Therefore, a true and accurate failure

Manuscript received June 4, 2013; accepted November 18, 2013. Date of publication December 3, 2013; date of current version March 4, 2014.

The author is with the Advanced Development Laboratory, Quantum Corporation, Irvine, CA 92617 USA (e-mail: Suayb.Arslan@Quantum.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TDMR.2013.2293491

1530-4388 © 2013 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

modeling shall be of great significance from a system design standpoint. For example, a generic hazard rate function $\lambda(x)$ and an associated non-homogenous Poisson model might be a better fit for predicting the real life disk failure trends. However, as more real life scenarios are incorporated with these improved mathematical methods for accuracy, they inexorably become complex. From a customer's perspective, short-hand closed-form expressions for predicting the system failure rates might be more useful for delivering performance figures about the system reliability.

In this study, an efficient MDS-parity based distributed disk array system is considered using general failure processes. One of the contributions of this paper is a set of useful closed-form expressions, derived by considering the whole lifespan of component drives based on the recent survey data on disk failures and time to failure probability distributions [8], i.e., without assuming constant component hazard rates. Some asymptotical results (the array size tends to infinity) shed light for the limiting behavior of RAID type systems. Those results might particularly be important for predicting what is achievable using MDS codes, as the scale of the coded storage systems grow for the management/maintenance of the so called "Big data." The paper also investigates the relationship between the aging and the redundancy used for data protection. Here, the efficiency of the distributed storage system comes rather from the efficient allocation strategy such that independent drive failure assumption is roughly correct for each component of the array, which are shared by different storage network nodes. It is further shown that the multidimensional array storage is offering a good tradeoff between complexity and performance which may otherwise be obtained by a large array of one dimensional RAID type systems at the expense of increased cost and complexity. Although the main objective of the paper is focused on the mean time to first failure, the expressions can be extended to mean time between failures and mean time to data loss performance metrics. However, derived expressions might either not be in simple form or expressible in a closed-form for an arbitrary hazard rate function $\lambda(x)$ and a repair rate function $\mu(x)$.

The remainder of this paper is organized as follows. In Section II, a brief introduction is given about the reliability theory basics as well as the drive failure statistics in real world. Moreover, the storage system details are summarized along with the assumptions used in this work. In Section III, main results of the paper are given based on arbitrary hazard rates using multidimensional arrays. This section starts with considering 1-D arrays and then generalizes the results for multidimensional arrays. Some of the numerical results and relevant examples are given in Section IV. Finally, a brief summary and conclusions follow in Section V. The proofs are included in Appendices A–C in order to highlight the main contributions of the paper.

II. Background and System Diagram

A. Reliability Theory

When a brand new product is put into service, it performs functional operations satisfactorily for a period of time, called *useful time* period, before eventually a failure occurs and the device is no longer able to respond to user requests. The observed time to failure (TTF) is a continuous random variable with a probability density function $f_{\mathrm{TTF}}(x)$, representing the lifetime of the product until the first failure. The failure probability of the device can be found using the cumulative distribution function (CDF) of TTF as follows,

$$F_{TTF}(x) = Pr\{TTF \leq x\} = \int_0^x f_{TTF}(y)dy,\ x > 0. \quad (1)$$

We can think of $F_{\mathrm{TTF}}(x)$ as an *unreliability* measure between time 0 and x. The reliability function $S(x)$ is therefore defined by,

$$S(x) \triangleq 1 - F_{TTF}(x) = \int_x^\infty f_{TTF}(y)dy. \quad (2)$$

In other words, reliability is the probability of having no failures before time x and is related to CDF of TFF through (2). Note that (2) implies that we have $f_{\mathrm{TTF}}(x) = -dS(x)/dx$. It may not be possible to estimate the distribution function of TFF directly from the available physical information. A useful function in clarifying the relationship between physical modes of failure and the probability distribution of TFF is known as the *hazard rate* function or *failure rate* function, denoted as $h_{\mathrm{TTF}}(x)$. This function is defined to be of the form

$$h_{TTF}(x) \triangleq \frac{f_{TTF}(x)}{S(x)} = -\frac{dS(x)}{S(x)dx}. \quad (3)$$

Solution of the first order ordinary differential equation (3) yields the relationship $h_{\mathrm{TTF}}(x) = -d(\ln(S(x)))/dx$ with the initial condition $S(0) = 1$. Note that knowing the hazard rate is equivalent to knowing the distribution. Mean time to failure (MTTF) is defined to be the expected value of the random variable TTF and is given by

$$MTTF \triangleq \mathbb{E}[TTF] = \int_0^\infty S(x)dx \Leftrightarrow \lim_{x\to\infty} xS(x) = 0 \quad (4)$$

where $\mathbb{E}[.]$ is the expectation operator. Note that (4) is true for distributions whose mean exists. For the rest of our discussion, the subscript TTF is dropped for notation simplicity and throughout the text, $h_{\mathrm{TTF}}(x)$ is alternatively denoted by $\lambda(x)$.

Annualized failure rate (AFR) is frequently used to estimate the failure probability of a device or a component after a full-time year use. In the conventional approach, time between failures are assumed to be independent and exponentially distributed with a constant rate λ. Therefore AFR is given by $AFR = 1 - S(x) = 1 - e^{-\lambda x}$, where $\lambda = 1/MTTF$ and x is the running time index in hours. MTTF is reported in hours and since there are 8760 hours in a year, $AFR = 1 - e^{-8760/MTTF}$. Typical numbers reported in disk vendor's specifications are $MTTF \approx 1$ million hours or 1.5 million hours. Since $8760/MTTF \ll 1$, $AFR \approx 8760/MTTF$. As mentioned before, such rough calculations and assumptions may not be representing how the drives behave in real world [8].

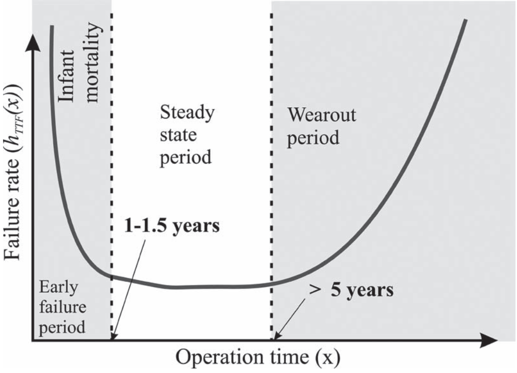


Fig. 1. Hazard rate pattern for hard disk drives as a function of operation time [13].

Clearly, one needs general but yet adequately simpler expressions to predict the lifetime trends of such systems, particularly for large-scale storage applications.

B. Drive Failures in Real World

It is shown in various research articles that average replacement rate of component disks is around twenty times much greater than are the theoretically predicted MTTF values i.e., predicted MTTF values are observed to be an underestimator [8], [13]. It is demonstrated that disk failure rates show a "bathtub" curve as shown in Fig. 1. Additionally, contrary to conventional homogenous stochastic models, hard disk replacement rates do not enter into steady state. After few years of use, drives (majority of which are disks) are observed to enter into wear-out period in which the failure rates steadily increase over time. Time between failures are shown to give much better fit with Weibull or gamma distributions instead of widely used exponential distribution [8]. There is a considerable amount of evidence that disk failures that are placed in the same batch show significant correlations, which is hard to quantify in a number of applications [14].

C. Efficient Storage System Summary

A series of parallel array of storage units (such as disk drives) is shown in Fig. 2(a). Drives are assumed to be manufactured identically and share the same failure/hazard rate function $\lambda(x)$. The $k_h \times k_v$ data matrix is encoded using two different block MDS codes in order to create a 2-D array. Horizontally, the parity information type-1 is computed using a $(n_h, k_h, t_h + 1)$ MDS code which can correct up to t_h erasures per block. It is due to the MDS property that it is the maximum number of erasures that a (n_h, k_h) block code can correct. Then, the computed parities are allocated to different disk units and occupy a fraction of the storage space of the disk array to protect the system against various types of system and disk failures. In addition to horizontal encoding, the parity information type-2 is computed using another $(n_v, k_v, t_v + 1)$ MDS code which can correct up to t_v erasures per vertical block. This encoding procedure can be performed repeatedly to protect larger dimensional data sets. The order of encoding does not matter as long as the code is a linear block code. A generalization of such an encoding scheme for three dimensional data is depicted in Fig. 2(b). Finally, encoded data units are allocated into the network storage nodes according to a genuine allocation policy that will keep

(I) the read and write process simple,

(II) the number of storage nodes needed to be accessed for the reconstruction of user/parity data minimum at a reasonable time complexity

(III) drives in the horizontal or vertical arrays (for 2-D array case) not shared by the same node of the distributed storage system.

These set of assumptions also help us make independent failure assumptions between the component drives of an array while in the mean time facilitate the rest of our analysis. For example, there is a class of MDS codes introduced in [15], for which the item (II) can be satisfied.

III. Disk Arrays With Individual Independent Arbitrary Hazard Rates

In the rest of our discussions, an allocation policy and a generic MDS code are assumed such that conditions (I), (II), and (III) are satisfied. Therefore, the rest of the discussion is based on the independent failure statistics assumption between respective drives of the storage array. A series of parallel arrays of disks is considered in which each array contains n disks or drives to store the encoded user data information. Let us use a common notation $(n, k, t + 1)$ where $t = n - k$ for the MDS code in order to make it general and applicable to each and every dimension to which erasure coding is applied.

A. A Horizontal System and Componentwise Reliability

Let us consider a 1-D array of storage units. Note that the results of this subsection can be applied to other arrays of different dimensions. This subsection starts with stating our main theorem below that bridges the relationship between redundancy and aging of MDS parity-based arrays.

Theorem 1: Hazard rate per data component of a horizontal system (consisting of n independent components k of which are data, each component having an arbitrary but the same failure rate of $\lambda(x)$), coded with a generic $(n, k, t + 1)$ block MDS code with rate $r = k/n$ ($t = n - k$ parity units) is given by

$$\mu_c(x, n, r) = \frac{\lambda(x)}{r}\left(1 - \frac{\psi_{t-1}(n-1, \lambda(x))}{\psi_t(n, \lambda(x))}\right) \tag{5}$$

where

$$\psi_t(z, n, \lambda(x)) \triangleq \sum_{i=0}^{t} \binom{n}{i}\binom{i}{z}(1 - R(x))^i R(x)^{n-i},$$

$R(x) = e^{-\int_0^x \lambda(y)dy}$ is the reliability of constituent components with hazard rate $\lambda(x)$ and $\psi_t(n, \lambda(x)) \triangleq \psi_t(0, n, \lambda(x))$ is the cumulative distribution function of the binomial distribution. Furthermore, the following inequality is satisfied for $0 \leq t \leq n - 1$,

$$\max\left\{0, \frac{1 - R(x)/r}{1 - R(x)}\right\} \leq \frac{\mu_c(x, n, r)}{\lambda(x)}. \tag{6}$$

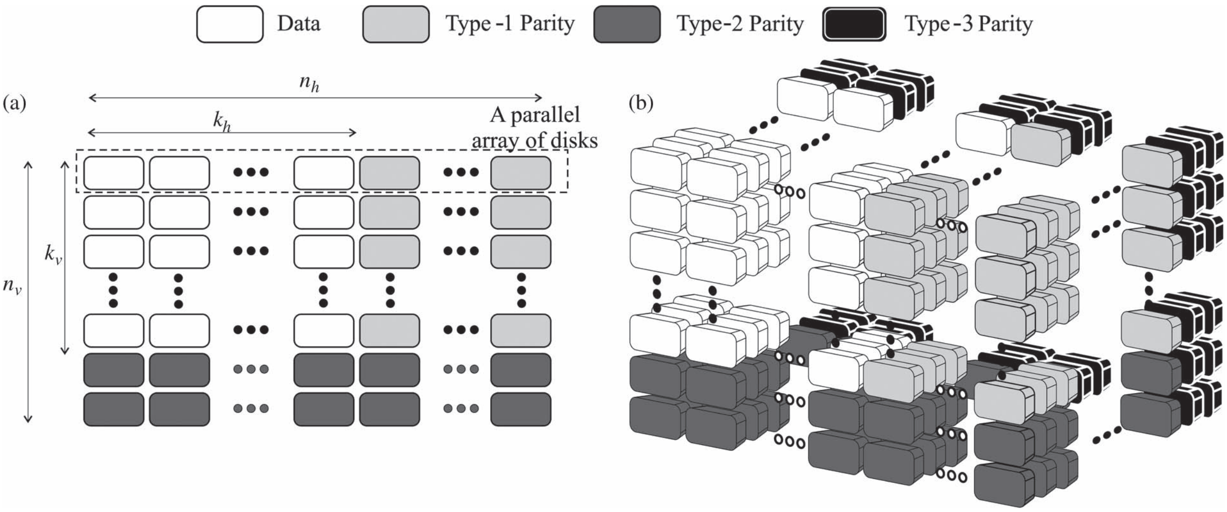


Fig. 2. Assume that the storage units are disks. (a) Parallel series of disk arrays that that makes up one disk matrix to be distributed over the network nodes. Each block represents a disk belonging to one of the parity types [13]. (b) A set of these disk matrices are used for storing large scale data using MDS encoding along each dimension.

Proof: See Appendix A.

Two cases shed some interesting light to this relationship. Consider the case with $t = 0$ and $r = 1$, i.e., no redundancy. In this case, since $\psi_{t-1}(n-1, \lambda(x)) = 0$ we have $\mu_c(x, n, 1) = \lambda(x)$ as expected. In other words, the hazard rate of the horizontal system per component is the same as the hazard rate of the constituent components when there is no redundancy. On the other extreme, we could have $t = n - 1$ and $r = 1/n$ using replication. In this case, the hazard rate per data component is given by the following *corollary*.

Corollary 2: Using a $(n, 1, n)$ block MDS code, known as repetition code, the hazard rate per data component is given by

$$\mu_c(x, n, 1/n) = \frac{n\lambda(x)R(x)\left(1 - R(x)\right)^{n-1}}{1 - \left(1 - R(x)\right)^n} \tag{7}$$

where $R(x) = e^{-\int_0^x \lambda(y)dy}$ is the reliability of constituent components with hazard rate $\lambda(x)$.

Proof: Let us set $t = n - 1$, we have

$$\begin{aligned}\psi_{n-1}\left(n, \lambda(x)\right) &= \psi_n\left(n, \lambda(x)\right) - \left(1 - R(x)\right)^n \\ &= 1 - \left(1 - R(x)\right)^n \qquad (8)\\ \psi_{n-2}\left(n-1, \lambda(x)\right) &= \psi_{n-1}\left(n-1, \lambda(x)\right) - \left(1 - R(x)\right)^{n-1} \\ &= 1 - \left(1 - R(x)\right)^{n-1}. \qquad (9)\end{aligned}$$

The result will follow through some algebraic manipulations by plugging (8) and (9) into (5). ■

A well known binary linear block MDS code is the parity code in which there is only one parity symbol i.e., $t = 1$ and $r = (n - 1)/n$. Following corollary characterizes the hazard rate of 1-D array using parity coding.

Corollary 3: Using a $(n, n - 1, 2)$ binary block MDS code, known as parity code, the hazard rate per data component of a horizontal block is given by

$$\mu_c(x, n, 1 - 1/n) = \frac{\lambda(x)n\left(1 - R(x)\right)}{n\left(1 - R(x)\right) + R(x)} \tag{10}$$

where $R(x) = e^{-\int_0^x \lambda(y)dy}$ is the reliability of constituent components with failure rate $\lambda(x)$.

Proof: We recognize that for $t = 1$ and,

$$\psi_0\left(n - 1, \lambda(x)\right) = R(x)^{n-1} \tag{11}$$

$$\psi_1\left(n, \lambda(x)\right) = R(x)^n + n\left(1 - R(x)\right)R(x)^{n-1}. \tag{12}$$

By plugging (11) and (12) into (5), with $r = 1 - 1/n$, we have

$$\begin{aligned}\mu_c(x, n, 1 - 1/n) &= \frac{\lambda(x)}{r}\left(1 - \frac{1}{R(x) + n\left(1 - R(x)\right)}\right) \qquad (13)\\ &= \frac{\lambda(x)n}{n-1}\left(\frac{(n-1)\left(1 - R(x)\right)}{R(x) + n\left(1 - R(x)\right)}\right) \qquad (14)\\ &= \frac{\lambda(x)n\left(1 - R(x)\right)}{n\left(1 - R(x)\right) + R(x)} \qquad (15)\end{aligned}$$

as desired. ■

In Theorem 1, if $R(x) \geq r$, the lower bound becomes zero, whereas if $R(x) < r$, the lower bound takes on a non-zero value. Let us define a system to be *componentwise reliable* (CR) if the hazard rate per drive component is zero or close to zero although the hazard rate of the whole system might be non-zero. Therefore, an interesting question is whether the lower bound of Theorem 1 is achievable for any real value of $R(x)$ and r as n grows large. This question will be explored next.

B. Asymptotical Hazard Rate Expressions

Regardless of the reliability distribution model used, as $x \to \infty$, the reliability of constituent components, $R(x)$ tends to zero. Therefore, one of the information Theorem 1 conveys is that for a fixed block length of n, if we let $R(x) \to 0$, we have the following lower bound,

$$\lim_{x \to \infty} \mu_c(x, n, r) \geq \lambda(x). \tag{16}$$

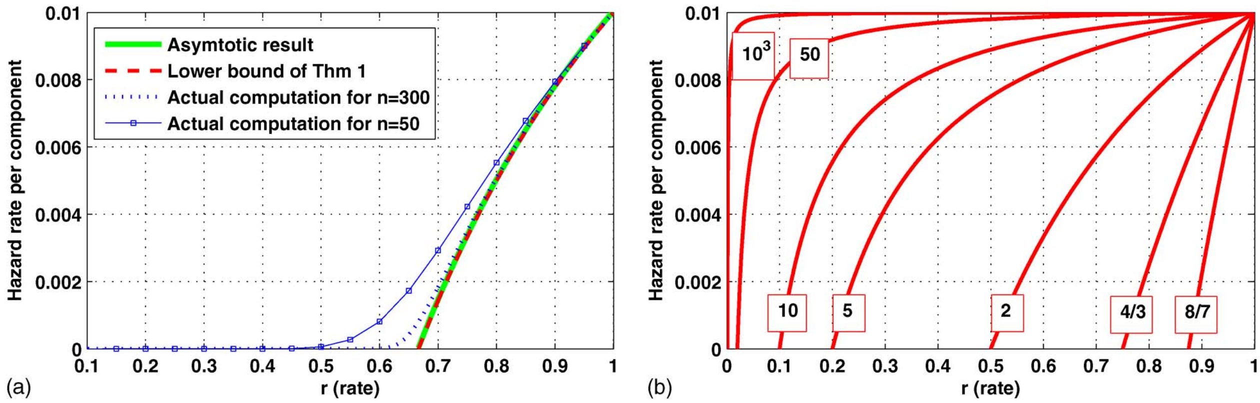


Fig. 3. (a) Asymptotic per component hazard rate when $R(a) = 1/q$ where $q = 3/2$ using an MDS code. (b) Asymptotically achievable bound as a function of rate r and q. As can be seen as q gets larger, the hazard rate per component tends to $\lambda(x)$ and become less dependent on rate. $\lambda(a) = 0.01$ is assumed.

Therefore, one might expect gains due to erasure coding in the infant mortality and useful time period, but not much in long term wear-out periods for 1-D arrays. As the number of component disks increases with the growing need for big data storage, the corresponding system reliability might be going down. It is of interest therefore to look at the asymptotic behavior, i.e., $n \to \infty$ of these reliability expressions at different times x.

Let us start with evaluating the asymptotic behavior $n \to \infty$ for a finite non-zero value of a such that $R(a) = 1/q$ for $q = 2, 3, \ldots$. In other words, for a given $\lambda(x)$, find a such that $\int_0^a \lambda(y)dy = \ln q$. For this special case, we have the following asymptotic result that shows the lower bound of Theorem 1 is achievable.

Theorem 4: Asymptotic hazard rate per data component of a horizontal system (as $n \to \infty$, each component having an arbitrary but the same failure rate of $\lambda(x)$), coded with a generic $(n, k, t+1)$ MDS block code with a fixed rate $r = k/n$ is given by

$$\mu_c(a,n,r) = \begin{cases} \frac{\lambda(a)(qr-1)}{r(q-1)} & \text{if } r \geq R(a) = \frac{1}{q} \\ 0 & \text{Otherwise} \end{cases} \qquad (17)$$

where a satisfies the relationship $\int_0^a \lambda(y)dy = \ln q$ for a given positive integer $q > 1$.

Proof: See Appendix B. The proof also conjectures that this theorem can be extended to any $q \in \mathbb{R}, q > 1$.

For $r \geq 1/q$, let us divide both the numerator and denumerator by qr and replace q with $1/R(a)$. We will have $\mu_c(a,n,r) = \lambda(a)((1-R(a))/r)/(1-R(a))$. Yet, this is the lower bound predicted by Theorem 1 evaluated at point $x = a$. Therefore, Theorem 4 proves that the lower bound of Theorem 1 is achievable for a countably infinite number of values of $R(x)$. We also conjecture that the lower bound of Theorem 1 is achievable for any value of $R(x)$ i.e., for any $q \in \mathbb{R}, q > 1$. Let us provide an example to support this conjecture by setting $q = 3/2$, a non-integer value, and $\lambda(arg \min_a\{|\int_0^a \lambda(y)dy - \ln q|\}) = 0.01$ is fixed for simplicity. We plot the asymptotic result, the lower bound due to Theorem 1 as well as the actual computation in Fig. 3(a). As can be seen asymptotic result of Theorem 4 achieves the lower bound of Theorem 1 for $r \geq 2/3$, below which we have a CR system if n is very large. However, if $n = 50$ or $n = 300$ we observe some performance loss and in order to obtain a CR system we must have $r \leq 0.12$ and $r \leq 0.46$, respectively.

So far, we have assumed that the size of the array n is increased for a given fixed value of $R(x)$. If $R(x) \to 0$ i.e., $q \to \infty$, we can see that the hazard rate per component of an MDS-protected array will converge to $\lambda(x)$. This is shown for a fixed value $\lambda = 0.01$ in Fig. 3(b). Yet, a general practice should be adaptively increasing the size n as $R(x)$ tends to zero i.e., as the reliability of components go down with time. The following theorem characterizes this scenario with the assumptions of an adaptive system: $\lim_{\substack{n\to\infty \\ x\to\infty}} nR(x) < \infty$ and $\lim_{\substack{n\to\infty \\ x\to 0}} n(1-R(x)) < \infty$ and shows that a CR system is possible for large-scale storage even if the component drives are in their wear-out period.

Theorem 5: if $\lim_{\substack{n\to\infty \\ x\to\infty}} nR(x) < \infty$ and $\lim_{\substack{n\to\infty \\ x\to 0}} n(1-R(x)) < \infty$, asymptotic hazard rate per data component of a horizontal system (as $n \to \infty$, each component having an arbitrary but the same failure rate of $\lambda(x)$), coded with a generic $(n, k, t+1)$ MDS block code with a fixed rate $r = k/n$ is given by

$$\lim_{\substack{n\to\infty \\ x\to a}} \mu_c(x,n,r) = \frac{\lambda(x)C(x,r,a)}{n} \qquad (18)$$

where

$$C(x,r,a) = \begin{cases} 1/r & \text{if } a = \infty \\ \frac{1-R(x)}{(R(x)-r)(2R(x)-r-1)} & \text{if } a = 0. \end{cases}$$

Proof: See Appendix C.

The amount of redundancy has a positive effect on the component hazard rate $\lambda(x)$ for this particular scenario. For a fixed n, the hazard rate for overall disk array was found to be scaling with $rn\lambda(x)$ if $n \to \infty$ first, then $R(x) \to 0$. On the other hand if $R(x) \to 0$ and $n \to \infty$ at the same time such that their product stays constant, this hazard rate converges to $\lambda(x)$, i.e., $k = rn$ times less than that of the fixed n case. Large block length improves the reliability performance at the expense of increased complexity. Note also that although the per component hazard rates might be tending to zero, the overall array hazard rate is nonzero.

The results of this subsection establishes an important relationship between the concept of CR and the rate r of the MDS code used. However, we assumed that the rate r is fixed through the whole lifespan of the storage system. Thus, it is easy to see that at some point in time x' we will have $r > R(x')$ and $\mu_c(x', n, r) \neq 0$ even if $n \to \infty$. Thus in order to obtain a CR system at all times, MDS codes with time varying rate $r(x)$ might be quite useful. In fact, there are asymptotically MDS codes called fountain codes that can be a perfect fit for this particular scenario [16]. Using such codes, rate can be adjusted on the fly such that $r(x) \leq R(x)$ is satisfied for all x if the condition of CR is strictly imposed on the design throughout the lifespan of the storage system. The design of such codes that also respects the set of assumptions (I), (II), and (III) for the particular application is beyond the scope of this paper.

C. Multidimensional Disk Arrays

Although the potential for a CR system is shown using large 1-D MDS-protected arrays, the implementation details and real life conditions make it impractical to achieve idealized performance benefits. Therefore, different directions must be taken for practical means such as multidimensional arrays using MDS codes. This is one of the natural ways to construct long blocks of many drives that can help us realize the asymptotical results derived in the previous subsection.

Previous section considered replaceable drive components in a 1-D horizontal structure and posed the question for any type of MDS code of rate r_h. Let us assume that we have a series of such parallel blocks of drives of hazard rate $k_h \mu_c(x, n_h, r_h)$ as shown in Fig. 2(a), generated by another MDS code of rate r_v. In other words, we have n_v 1-D arrays of size n_h drives each, such that if more than $n_v(1 - r_v)$ blocks fail, it will lead to the whole system failure. This is due to the MDS property of the erasure correction coding. Furthermore, we assume inter-block failure independence and if a horizontal block fails, all the constituent disks are assumed to be failed. For a general case, this type of decoding procedure corresponds to the failure of T-D disk array, if at least one (T-1)-D disk array fails. More complicated decoding procedures can be employed for better performance at the expense of increased implementation complexity.

Let $\mu_c(x, n_h n_v, r_h r_v)$ be the hazard rate per data component of the 2-D disk array. Using the result of Theorem 1, we shall obtain

$$\mu_c(x, n_h n_v, r_h r_v) = \frac{\mu_c(x, n_h, r_h)}{k_h r_v} \times \left(1 - \frac{\psi_{(1-r_v)n_v - 1}\left(n_v - 1, k_h \mu_c(x, n_h, r_h)\right)}{\psi_{(1-r_v)n_v}\left(n_v, k_h \mu_c(x, n_h, r_h)\right)}\right). \quad (19)$$

The result follows from Theorem 1 by replacing $\lambda(x)$ with $k_h \mu_c(x, n_h, r_h)$ and n with n_v. Finally, the system failure rate is divided by the total number of data disks in a horizontal array. This expression is indeed a special case of the following more general result on T-D disk array system encoded with a set of MDS block codes with parameters $\{(n_1, r_1), (n_2, r_2), \ldots, (n_T, r_T)\}$. Note that 3-D case is shown in Fig. 2(b) and larger dimensional generalizations are possible yet are hard to visualize.

Theorem 6: For T-D MDS-protected system of drives or disks, we have the following recursive formulation that establishes the relationship in terms of the per component rate of (T-1)-D array,

$$\mu_c(x, n_{1,T}, r_{1,T}) = \frac{\mu_c(x, n_{1,T-1}, r_{1,T-1})}{r_{1,T} n_{1,T-1}} \quad (20)$$

$$\times \left(1 - \frac{\psi_{(1-r_T)n_T - 1}(n_T - 1, k_{1,T-1}\mu_c(x, n_{1,T-1}, r_{1,T-1}))}{\psi_{(1-r_T)n_T}(n_T, k_{1,T-1}\mu_c(x, n_{1,T-1}, r_{1,T-1}))}\right) \quad (21)$$

where $n_{1,s} \triangleq \prod_{i=1}^{s} n_i$, $r_{1,s} \triangleq \prod_{i=1}^{s} r_i$ and $k_{1,s} = r_{1,s} n_{1,s}$.

Proof: (sketch) Proof follows from Theorem 1 by replacing $\lambda(x)$ with $k_{1,T-1}\mu_c(x, n_{1,T-1}, r_{1,T-1})$ in which $\mu_c(x, n_{1,T-1}, r_{1,T-1})$ is the hazard rate per component for (T-1)-D parity protected system and $k_{1,T-1}$ is the number of data component disks. In this case, the result of Theorem 1 can be applied to compute the hazard rate of a disk hyperplane of dimension T-1. In order to find the hazard rate per component, we divide the overall hazard rate function by the number of components $k_{1,T-1}$. We eventually obtain the result using the fact $r_{1,T-1} r_T = r_{1,T}$. ■

IV. Examples and Numerical Results

In this section, results will be provided for some of the special cases for finite block lengths so that a comparison can be made with asymptotical results. Reliability of multidimensional arrays will be compared in terms of component as well as array level hazard rates.

1) Constant Hazard Rate Components With $r = 1/n$: Consider a parallel block and the non-aging components with a constant and identical rate i.e., $\lambda(x) = \lambda$. Using Corollary 2 and $R(x) = e^{-\lambda x}$ for constant rate $\lambda(x) = \lambda$, as derived in [17], we have

$$\mu_c(x, n, r) = \frac{n\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{n-1}}{1 - (1 - e^{-\lambda x})^n}. \quad (22)$$

Note that this constant failure rate assumption was originally used by manufactures to predict the failure trends. Equation (22) can be approximated as $\mu_c(x, n, r) \approx n\lambda^n x^{n-1}$ when $x \ll 1/\lambda$. As $x \to \infty$, the hazard rate converges to λ, achieving the lower bound of Theorem 1. In its early use, the system failure rate grows as a power function of age which is known as the Weibull law. This means that using more redundancy within the block triggers aging although the constituent parts are non-aging components.

2) Non-Constant Hazard Rate Components Using Arbitrary r: Let us consider a general form of $\lambda(x)$ that is in bathtub shape using a composite distribution model[1] given by

$$\lambda(x) = \begin{cases} \frac{\beta_1 t^{\beta_1 - 1}}{\theta_1^{\beta}} & \text{if } 0 < t \leq t_1 \\ \frac{\beta_2 t^{\beta_2 - 1}}{\theta_2^{\beta_2}} & \text{if } t_1 < t \leq t_2 \\ \frac{\beta_3 t^{\beta_3 - 1}}{\theta_3^{\beta_3}} & \text{if } t > t_2 \end{cases} \quad (23)$$

[1] A Weibull hazard function is used to model three different periods of failure processes by appropriately choosing the parameters of the distribution.

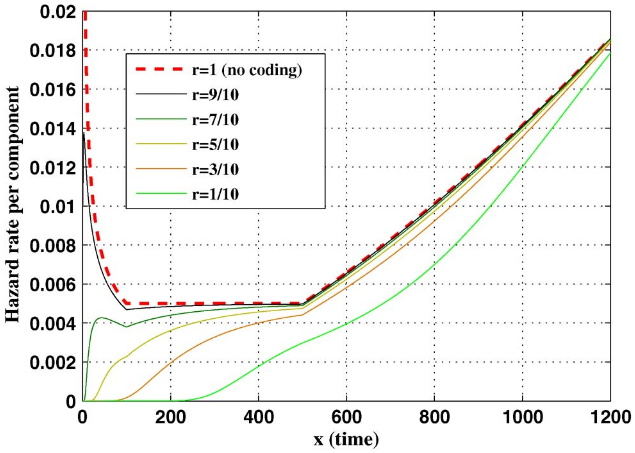


Fig. 4. Hazard rate function per component for 1-D disk array system as a function of time with and without coding. $\lambda(x)$ is assumed to be a simple bathtub curve obtained by using a composite distribution based on Weibull models. Let us assume $n = 100$ disks per array.

with the corresponding reliability function,

$$R(x) = \begin{cases} e^{-\left(\frac{x}{\theta_1}\right)^{\beta_1 - 1}} & \text{if } 0 < t \leq t_1 \\ e^{-\left(\frac{x-t_1}{\theta_2}\right)^{\beta_2 - 1} - \left(\frac{t_1}{\theta_1}\right)^{\beta_1 - 1}} & \text{if } t_1 < t \leq t_2 \\ e^{-\left(\frac{x-t_2}{\theta_3}\right)^{\beta_3 - 1} - \left(\frac{t_2-t_1}{\theta_2}\right)^{\beta_2 - 1} - \left(\frac{t_1}{\theta_1}\right)^{\beta_1 - 1}} & \text{if } t > t_2. \end{cases} \tag{24}$$

For useful life period (random failure process) between time t_1 and t_2, let us set $\beta_2 = 1$ and $\theta_2 = 200$. For the early life (infancy), we use $\beta_1 = 0.5$ and $\theta = 100$ to model the decreasing failure rates. In order to model the wear-out period, let us set $\beta_3 = 2.5$ and $\theta_3 = 500$. An example is shown in Fig. 4 for an array of disks ($n = 100$) using MDS codes with different rates. As predicted by asymptotical expressions, for $x \to 0$ we have $\mu_c(x, 100, r) \to 0$ and for $x \to \infty$ we have $\mu_c(x, 100, r) \to \lambda(x)$. This figure also suggests that there is a key number of parities such that the useful life time period can be widened i.e., we have constant failure rates for longer period of time with coding. Another interesting observation is that coded system improves the wear-out period only if exceeding number of parities are used i.e. a rate of 1/10 gives us a reasonable improvement although the aging can be greatly reduced at early life period for each component disk.

In the next set of results, let us compare 1-D array of $n = 300$ disks in which half of the disks are dedicated to parity to a 2-D array of disks of the same size and relative redundancy, encoded with (25, 15, 11) and (12, 10, 3) MDS codes both in horizontal and vertical directions, respectively. Let us focus on array hazard rates i.e., data loss rate rather than individual disk hazard rates and assume independence. The results are presented in log-log scale in Fig. 5. As can be seen, using a 2-D coding structure the data loss rate can greatly be lessened. In fact, since the block length and the rate of the component codes of the 2-D MDS code are reduced, some performance loss is observed at the infancy period. However, 2-D structure might be easier to implement as the MDS codes of shorter length and larger rate are utilized.

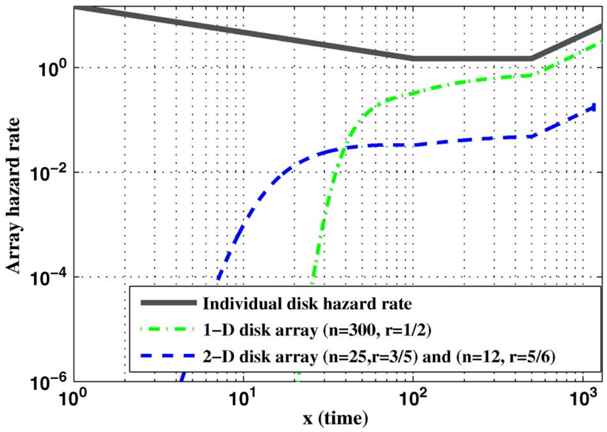


Fig. 5. Array hazard rate function using different MDS codes and structures. $\lambda(x)$ is assumed to be a simple bathtub curve obtained by using a composite distribution based on Weibull assumption. Number of disks in each system is 300 and the total rate of each MDS code is 1/2.

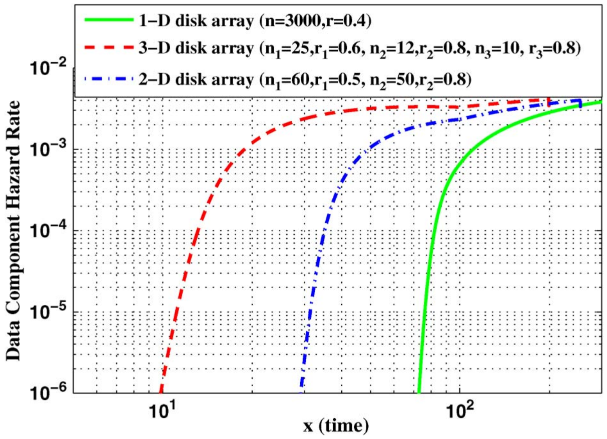


Fig. 6. Component hazard rate function using multidimensional MDS codes and structures. $\lambda(x)$ is assumed to be a simple bathtub curve obtained by using a composite distribution based on Weibull models. Number of disks in each system is 3000 and the total rate of each MDS code is 0.4.

Let us look at the individual data component disk hazard rates for a 3-D structure with an overall rate 0.4 and $n =$ 3000 disks. Fig. 6 shows the data component disk hazard rates for 1-D, 2-D and 3-D disk array structures of the same size. Component MDS codes for 2-D structure have the parameters (60, 30, 31) and (50, 40, 11) whereas for 3-D structure, they have the parameters (25,15,11), (12, 10, 3), and (10, 8, 3). As can be observed, although the array hazard rates show better performance with multi-dimensional MDS structures, the component hazard rates are worse compared to that of 1-D disk array. This is mainly due to 1-D MDS-protected system performs better (in fact it comes close to the performance lower bound) than multiple short block length MDS codes of the same rate. However, the performance gain of multidimensional disk structures is rather in terms of low complexity due to using short block length and high rate MDS codes. In addition, if the array of disks fail for 1-D structure, the whole system fails and data is lost. On the other hand, multidimensional structures have many arrays of shorter length and it is low probability to lose all of them at once. It is not hard to see that data component disk hazard rates of multidimensional structures (2-D and 3-D arrays in our case) are close to that of 1-D array, particularly for useful and wear-out time periods. This means that the lower bounds can be achieved using multidimensional structures for a large fraction of time of a drive's lifespan. Finally, we note that our arguments are based on a conventional decoding algorithm used for product codes, more advanced algorithms might improve the over all decoding performance.

V. Conclusion

Generalized expressions are given for disk arrays that are protected by MDS-parities. A brief analysis of the interaction is also presented between redundancy and aging of MDS-parity based disk array systems in a distributed storage scenario. The relationship between redundancy level and aging is demonstrated using general formulations and accurate distributions that is more reflective of the real life failure scenarios. Asymptotic results show that performance lower bounds are achievable with large-scale storage networks as long as independence is assumed among the component failures. Although neat compact form expressions may not exist for some of the derivations, numerical results provide some intuition about the behavior of such disk arrays under independent failure modes. Results are extended to include multidimensional disk arrays to show that there might be practical ways to get close to predicted performance lower bounds for the component hazard rates. We have not specified any particular code to keep the derived expressions more general. One property of these codes was their optimality, characterized by the MDS feature. However, other near-MDS codes and many variations of such can be used in storage systems due to implementation and the expressions derived in this paper shall still be applicable. Discussions regarding a particular code choice without the MDS property and the associated reliability expressions are beyond the scope of this paper.

Appendix A
Proof of Theorem 1

Let us begin this section with the following lemma.

Lemma 7: The function $\psi_t(z, n, \lambda(x))$ satisfies the following relationship for any integer z, satisfying $0 \leq z \leq n$

$$\frac{\psi_t(z, n, \lambda(x))}{\psi_{t-z}(n-z, \lambda(x))} = \binom{n}{z}(1-R(x))^z \tag{25}$$

where $\psi_t(n, \lambda(x)) \triangleq \psi_t(0, n, \lambda(x))$ is the cumulative distribution function of the binomial distribution.

Proof: First note that for $z > i$, we have the convention $\binom{i}{z} = 0$. Therefore we rewrite the expression for $\psi_t(z, n, \lambda(x))$,

$$= \sum_{i=z}^{t} \binom{n}{i}\binom{i}{z}(1-R(x))^i R(x)^{n-i} \tag{26}$$

$$= \sum_{i=z}^{t} \binom{n-z}{i-z}\binom{n}{z}(1-R(x))^i R(x)^{n-i} \tag{27}$$

$$= \binom{n}{z}(1-R(x))^z \sum_{i=z}^{t} \binom{n-z}{i-z}(1-R(x))^{i-z} R(x)^{n-i} \tag{28}$$

$$= \binom{n}{z}(1-R(x))^z \sum_{j=0}^{t-z} \binom{n-z}{j}(1-R(x))^{j} R(x)^{n-z-j} \tag{29}$$

$$= \binom{n}{z}(1-R(x))^z \psi_{t-z}(n-z, \lambda(x)) \tag{30}$$

from which the result follows. Note that we make the change of variables $j = i - z$ in (29) and the (27) follows from binomial coefficient identity $\binom{n}{i}\binom{i}{z} = \binom{n-z}{i-z}\binom{n}{z}$. ■

After establishing a useful lemma, let us give the proof of *Theorem 1*. It is clear that a horizontal block failure will occur only if $t+1$ or more drives fail in the horizontal block of size n disks.[2] Due to independence assumption, the reliability of such a block is given by

$$S(x) = \sum_{i=0}^{t} \binom{n}{i}(1-R(x))^i R(x)^{(n-i)} = \psi_t(n, \lambda(x)). \tag{31}$$

Let us find the probability density function of the time between component failures. This is given by

$$f_X(x) = -S'(x) = -\frac{dS(x)}{dx} \tag{32}$$

$$\begin{aligned} &= \sum_{i=0}^{t} \binom{n}{i} \left\{ \frac{iR'(x)}{1-R(x)}(1-R(x))^i R(x)^{(n-i)} \right. \\ &\qquad - \frac{(n-i)R'(x)}{R(x)}(1-R(x))^i \\ &\qquad \left. \times R(x)^{(n-i)} \right\} \end{aligned} \tag{33}$$

$$\begin{aligned} &= -n\frac{R'(x)}{R(x)}\psi_t(n, \lambda(x)) \\ &\quad + \frac{R'(x)}{R(x)}\frac{1}{1-R(x)}\psi_t(1, n, \lambda(x)) \end{aligned} \tag{34}$$

$$= \lambda(x)n\psi_t(n, \lambda(x)) - \frac{\lambda(x)\psi_t(1, n, \lambda(x))}{1-R(x)} \tag{35}$$

where we used the fact that $R'(x)/R(x) = -\lambda(x)$. Thus, finally from (3), we compute the total hazard rate for all k data components (a series of k data storage units) as follows,[3]

$$\begin{aligned} k\mu_c(x, n, r) &= \frac{f_X(x)}{S(x)} = \frac{f_X(x)}{\psi_t(n, \lambda(x))} \\ &= \lambda(x)\left(n - \frac{\psi_t(1, n, \lambda(x))/\psi_t(n, \lambda(x))}{1-R(x)}\right). \end{aligned} \tag{36}$$

If we use the result of Lemma 7 with $z = 1$, i.e.,

$$\psi_t(1, n, \lambda(x)) = n(1-R(x))\psi_{t-1}(n-1, \lambda(x)) \tag{37}$$

and (36) then we arrive at (5).

As for the lower bound, we observe the following relationship due to $t \geq i$,

$$t\psi_t(0, n, \lambda(x)) \geq \psi_t(1, n, \lambda(x)). \tag{38}$$

[2] Since the controller of disk array systems can identify which disks are failed, MDS codes are used to correct erasures.

[3] We note that the horizontal system failure rate is always equal to the sum of the component failure rates, regardless of the distributions used to describe the components. In other words, let $\lambda_H(x)$ be the failure rate of a horizontal system that consists of N components. If the component failure rates are characterized by the set of hazard rates $\{\lambda_i(x)\}_{i=1}^{N}$, it is not hard to show that $\lambda_H(x) = \sum_{i=1}^{N} \lambda_i(x)$. This result is used implicitly throughout the paper to find the per component hazard rates of multidimensional coded storage systems.

Using above equation and (37), we can proceed as follows,

$$\mu_c(x,n,r) \geq \frac{\lambda(x)}{r}\left(1 - \frac{t}{n\left(1-R(x)\right)}\right) \tag{39}$$
$$= \frac{\lambda(x)}{r}\left(\frac{r-R(x)}{1-R(x)}\right) \tag{40}$$

from which the lower bound follows. Notice that if $R(x) > r$, then this lower bound takes on negative values. Therefore, the maximum operator is introduced to make the lower bound non-negative. ■

Appendix B
Proof of Theorem 4

Let $R(a) = 1/q$ for some $a > 0, q > 1$, we have

$$\psi_{t-1}\left(n-1, \lambda(a)\right) = \sum_{i=0}^{t-1}\binom{n-1}{i}\left(1-\frac{1}{q}\right)^{i}\left(\frac{1}{q}\right)^{n-1-i}$$
$$= q^{-n+1}\sum_{i=0}^{t-1}\binom{n-1}{i}\left(\frac{q-1}{q}\right)^{i}\left(\frac{1}{q}\right)^{-i}$$
$$= q^{-n+1}\sum_{i=0}^{t-1}\binom{n-1}{i}(q-1)^{i} \tag{41}$$

and similarly,

$$\psi_t\left(n, \lambda(a)\right) = q^{-n}\sum_{i=0}^{t}\binom{n}{i}(q-1)^{i}. \tag{42}$$

Using an asymptotic result from coding theory that in a q-ary n dimensional linear space $\mathbb{F}_q^n$, the volume of Hamming spheres (balls) of radius t can be bounded for large n and $t/n = 1 - r \leq 1 - 1/q$ i.e., $r \geq 1/q$ by

$$q^{(h_q(1-r)-o(1))n} \leq \sum_{i=0}^{t}\binom{n}{i}(q-1)^{i} \leq q^{h_q(1-r)n} \tag{43}$$

where $h_q(p)$ is the q-ary entropy function given by

$$h_q(p) \triangleq p\log_q(q-1) + p\log_q\left(\frac{1}{p}\right) + (1-p)\log_q\left(\frac{1}{1-p}\right) \tag{44}$$

and $o(1) \to 0$ as $n \to \infty$. In the context of coding theory, q is usually an integer representing the size of the alphabet over which the code is defined. Here in our case, it is not to hard to show that (43) is valid for any value of $q \in \mathbb{R}$ as long as $qr \geq 1$. Finally, we observe that

$$q^{h_q(1-\frac{rn}{n-1})(n-1)-o(n)-h_q(1-r)n+1}$$
$$\leq \frac{\psi_{t-1}\left(n-1,\lambda(a)\right)}{\psi_t\left(n,\lambda(a)\right)}$$
$$\leq q^{h_q(1-\frac{rn}{n-1})(n-1)-h_q(1-r)n+o(n)+1} \tag{45}$$

and $\exists\epsilon > 0$ such that the following assures the convergence for large n,

$$Pr\left\{\left|\frac{\psi_{t-1}\left(n-1,\lambda(a)\right)}{\psi_t\left(n,\lambda(a)\right)} - q^{(n-1)h_q\left(1-\frac{rn}{n-1}\right)-nh_q(1-r)+1}\right| > \epsilon\right\} = 0. \tag{46}$$

Let us use the definition for $h_q(p)$ to expand our expression along with the asymptotical results that $\lim_{n\to\infty}\log_q(1-rn/n-1) = \log_q(1-r)$ and $\lim_{n\to\infty}\log_q(rn/n-1) = \log_q(r)$

$$\log_q\left(q^{h_q\left(1-\frac{rn}{n-1}\right)(n-1)-h_q(1-r)n+1}\right)$$
$$= (n-1)h_q\left(1-\frac{rn}{n-1}\right) - nh_q(1-r) + 1 \tag{47}$$

where

$$(n-1)h_q\left(1-\frac{rn}{n-1}\right)$$
$$= (n-1)\left[\left(1-\frac{rn}{n-1}\right)\log_q(q-1) - \left(1-\frac{rn}{n-1}\right)\log_q(1-r) - \left(\frac{rn}{n-1}\right)\log_q(r)\right] \tag{48}$$
$$= (n-1-rn)\log_q(q-1) - (n-1-rn)\log_q(1-r) - rn\log_q(r) \tag{49}$$

and similarly,

$$nh_q(1-r) = (n-rn)\log_q(q-1) - (n-rn)\log_q(1-r) - rn\log_q(r). \tag{50}$$

Finally, let us use (49) and (50) to obtain,

$$(n-1)h_q\left(1-\frac{rn}{n-1}\right) - nh_q(1-r) + 1$$
$$= 1 - \log_q(q-1) + \log_q(1-r) \tag{51}$$
$$= \log_q\left(\frac{q(1-r)}{q-1}\right). \tag{52}$$

This result justifies that, for large enough n, we have the following convergence

$$q^{(n-1)h_q\left(1-\frac{rn}{n-1}\right)-nh_q(1-r)+1} \to \frac{q(1-r)}{q-1} \tag{53}$$

which completes the proof for $r \geq 1/q$. For $r \leq 1/q$, we have

$$\frac{q(1-r)}{q-1} \geq 1. \tag{54}$$

On the other hand, It is easy to see that for $t \leq n-1$

$$\sum_{i=0}^{t}\binom{n-1}{i}(q-1)^{i+1}$$
$$\geq \sum_{i=0}^{t-1}\binom{n-1}{i}(q-1)^{i+1} \tag{55}$$
$$= \sum_{i=0}^{t}\binom{n-1}{i-1}(q-1)^{i} \tag{56}$$
$$= \sum_{i=0}^{t}\left[\binom{n}{i} - \binom{n-1}{i}\right](q-1)^{i} \tag{57}$$

from which we obtain,

$$q\sum_{i=0}^{t}\binom{n-1}{i}(q-1)^i \geq \sum_{i=0}^{t}\binom{n}{i}(q-1)^i. \tag{58}$$

Using a similar argument, we can show that

$$\frac{\psi_{t-1}\left(n-1,\lambda(a)\right)}{\psi_t\left(n,\lambda(a)\right)} = \frac{q}{q-1}\left(1-\frac{\sum_{i=0}^{t}\binom{n-1}{i}(q-1)^i}{\sum_{i=0}^{t}\binom{n}{i}(q-1)^i}\right) \tag{59}$$

and using (58), we obtain

$$\frac{\psi_{t-1}\left(n-1,\lambda(a)\right)}{\psi_t\left(n,\lambda(a)\right)} \leq 1. \tag{60}$$

Combining (54) and (60), we have $\psi_{t-1}(n-1,\lambda(a))/\psi_t(n,\lambda(a)) = 1$ for $n \to \infty$. Finally, this implies $\mu_c(a,n,r) \to 0$ as $n \to \infty$ if $r \leq 1/q$. ■

Appendix C
Proof of Theorem 5

Before proving Theorem 5, let us first prove the following useful lemmas.

Lemma 8: The ratio of an incomplete Gamma function to a complete Gamma function satisfies the following relationship,

$$1-\frac{b}{a-b-1} < \frac{\Gamma(a,b)}{\Gamma(a)} = \frac{\Gamma(a,b)}{(a-1)!} < 1 \tag{61}$$

where $\Gamma(a,b) = \int_b^\infty t^{a-1}e^{-t}dt$ is the incomplete Gamma function.

Proof: Let us explore the difference,

$$\Gamma(a)-\Gamma(a,b) = \int_0^b t^{a-1}e^{-t}dt < \int_0^b b^{a-1}e^{-b}dt \tag{62}$$

$$= b^a e^{-b}$$

$$= \frac{b}{a-b-1}\int_b^{a-1} b^{a-1}e^{-b}dt$$

$$< \frac{b}{a-b-1}\int_b^{a-1} t^{a-1}e^{-t}dt \tag{63}$$

$$< \frac{b}{a-b-1}\int_0^{\infty} t^{a-1}e^{-t}dt \tag{64}$$

$$= \frac{b}{a-b-1}\Gamma(a) \tag{65}$$

which establishes the lower bound. The upper bound follows from the definition of incomplete beta function. Note that the integrand $t^{a-1}e^{-t}$ achieves its maximum at $t = a-1$ and for $0 < t < a-1$, it is increasing. Thus, for $0 < t < b$, we have the inequality (62) and for $b < t < a-1$ we have the inequality (63). ■

Lemma 8 indicates that for a fixed $b > 0$ and $a \to \infty$, $\Gamma(a,b) \to \Gamma(a)$. In addition, for a fixed $b > 0$ and $a \gg b$, we have the following approximation

$$\frac{\Gamma(a,b)}{\Gamma(a)} \approx \frac{a-2b-1}{a-b-1}. \tag{66}$$

Lemma 9: As $n \to \infty$ and $R(x) \to 0$ while satisfying $\lim_{\substack{n\to\infty\\x\to\infty}} nR(x) < \infty$, we have the following convergence

$$\psi_t\left(n,\lambda(x)\right) \longrightarrow 1-\frac{\Gamma\left(k,nR(x)\right)}{\Gamma(k)} \approx \frac{nR(x)}{k-nR(x)-1} \tag{67}$$

and as $n \to \infty$ and $x \to 0$ while satisfying $\lim_{\substack{n\to\infty\\x\to 0}} n(1-R(x)) < \infty$, we have the following convergence

$$\psi_t\left(n,\lambda(x)\right) \longrightarrow \frac{\Gamma\left(t+1,n\left(1-R(x)\right)\right)}{\Gamma(t+1)} \approx \frac{2R(x)-r-1}{R(x)-r} \tag{68}$$

where $R(x) = e^{-\int_0^x \lambda(y)dy}$.

Proof: First note the following relationship,

$$\psi_t\left(n,\lambda(x)\right) = \sum_{i=0}^{t}\binom{n}{i}\left(1-R(x)\right)^i R(x)^{n-i}$$

$$= \sum_{j=n-t}^{n}\binom{n}{j}R(x)^j\left(1-R(x)\right)^{n-j}$$

$$= 1-\sum_{j=0}^{n-t-1}\binom{n}{j}R(x)^j\left(1-R(x)\right)^{n-j}. \tag{69}$$

If $\lim_{\substack{n\to\infty\\x\to\infty}} nR(x) < \infty$, then the asymptotical convergence of binomial distribution to Poisson distribution yields

$$\binom{n}{j}R(x)^j\left(1-R(x)\right)^{n-j} \to \frac{n^jR(x)^je^{-nR(x)}}{j!}. \tag{70}$$

Thus, for sufficiently large n, we have

$$\psi_t\left(n,\lambda(x)\right) = 1-\sum_{j=0}^{n-t-1}\frac{n^jR(x)^je^{-nR(x)}}{j!} \tag{71}$$

$$= 1-\frac{\Gamma\left(n-t,nR(x)\right)}{(n-t-1)!} \tag{72}$$

$$= 1-\frac{\Gamma\left(k,nR(x)\right)}{\Gamma(k)} \tag{73}$$

and using the approximation (66) with $a = k$ and $b = nR(x)$, the (67) follows. Note that if $n \to \infty$, then $k = nr \to \infty$. Similarly, using (69) and following the same line of proof, we can obtain (68) ■

Next, we give the proof of *Theorem 5*. First, consider $n \to \infty$ and $R(x) \to 0$ while satisfying $\lim_{\substack{n\to\infty\\x\to\infty}} nR(x) < \infty$. Using the results of Lemma 8 and Lemma 9, we can approximate the limiting ratio,

$$1-\frac{\psi_{t-1}\left(n-1,\lambda(x)\right)}{\psi_t\left(n,\lambda(x)\right)} \to 1-\frac{(n-1)R(x)}{nR(x)} \times \frac{nr-nR(x)-1}{nr-(n-1)R(x)-1}$$

$$= \frac{nr-1}{n\left(nr-(n-1)R(x)-1\right)} \tag{74}$$

$$\approx \frac{1}{n}. \tag{75}$$

Therefore, we have $k\mu_c(x,n,r) \rightarrow k\lambda(x)/nr = \lambda(x)$ which establishes what is asserted.

Similarly, let us consider $n \rightarrow \infty$ and $x \rightarrow 0$ while satisfying $\lim_{\substack{n\rightarrow\infty \\ x\rightarrow 0}} n(1-R(x)) < \infty$. Let $b = (n-1)(1-R(x))$ and use lemma 8, Lemma 9 and the approximation (66). Employing some algebraic manipulations, we can obtain

$$\begin{aligned} n\left(1-\frac{\psi_{t-1}\left(n-1,\lambda(x)\right)}{\psi_t\left(n,\lambda(x)\right)}\right) &\approx n\left(1-\frac{1-b/(t-b-1)}{1-\frac{b+1-R(x)}{t-b+R(x)-1}}\right) &(76)\\ &= \frac{R(x)(t-1)-t+b+1}{(t-b-1)\left(t-2b+2R(x)-2\right)}. &(77)\end{aligned}$$

Let us divide both the numerator and denominator by n^2. Letting $n\rightarrow\infty$, we can approximate the limiting ratio as follows

$$n\left(1-\frac{\psi_{t-1}\left(n-1,\lambda(x)\right)}{\psi_t\left(n,\lambda(x)\right)}\right)\rightarrow\frac{r\left(1-R(x)\right)}{\left(R(x)-r\right)\left(2R(x)-r-1\right).}$$

Therefore using *Theorem 1*, we have

$$\begin{aligned} k\mu_c(x,n,r) &= n\lambda(x)\left(1-\frac{\psi_{t-1}\left(n-1,\lambda(x)\right)}{\psi_t\left(n,\lambda(x)\right)}\right) &(78)\\ &\rightarrow\frac{\lambda(x)r\left(1-R(x)\right)}{\left(R(x)-r\right)\left(2R(x)-r-1\right)} &(79)\end{aligned}$$

which completes the proof. ■

References

[1] D. A. Patterson, G. A. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in *Proc. ACM Conf. Manag. Data (SIGMOD)*, Chicago, IL, USA, Jun. 1988, pp. 109–116.

[2] Y. Kim, S. Oral, D. Dillow, F. Wang, F. S. Poole, and G. Shipman, "An empirical study of redundant array of independent Solid-State Drives (RAIS)," Natl. Center Comput. Sci., Oak Ridge Natl. Lab., Oak Ridge, TN, USA, Tech. Rep. (ORNL/TM-2010/61), 2010.

[3] *Specification of Hard Disk Drive Reliability*, IDEMA Standard R2-98, 1998.

[4] M. Malhotra, "Reliability analysis of redundant arrays of inexpensive disks," *J. Parall. Distrib. Comput.*, vol. 17, no. 1/2, pp. 146–151, Jan. 1993.

[5] K. Salem and H. Garcia-Molina, "Disk striping," in *Proc. 2nd IEEE Int. Conf. Data Eng.*, 1986, pp. 336–342.

[6] R. E. Blahut, *Algebraic Codes for Data Transmission*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[7] G. A. Gibson and D. A. Patterson, "Designing disk arrays for high data reliability," *J. Parall. Distrib. Comput.*, vol. 17, no. 1/2, pp. 4–27, Jan./Feb. 1993.

[8] B. Schroeder and G. A. Gibson, "Disk failures in the real world: What does an MTTF of 1 000 000 hours mean to you?" in *Proc. 5th USENIX Conf. FAST*, Feb. 2007, pp. 1–16.

[9] S. Shah and J. G. Elerath, "Reliability analysis of disk drive failure mechanisms," in *Proc. Annu. Reliab. Maintainability Symp.*, 2005, pp. 226–231.

[10] J. Elerath and M. Pecht, "Enhanced reliability modeling of RAID storage systems," in *Proc. Int. Conf. DSN*, Edinburgh, U.K., Jun. 2007, pp. 175–184.

[11] H. E. Ascher, "A set-of-numbers is NOT a DataSet," *IEEE Trans. Reliab.*, vol. 48, no. 2, pp. 135–140, Jun. 1999.

[12] T. J. E. Schwarz, Q. Xin, E. L. Miller, D. D. E. Long, A. Hospodor, and S. Ng, "Disk scrubbing in large archival storage systems," in *Proc. IEEE Comput. Soc. Symp. MASCOTS*, 2004, pp. 409–418.

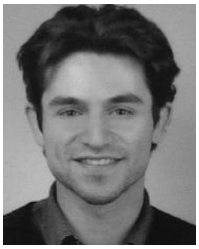
[13] J. Yang and F.-B. Sun, "A comprehensive review of hard-disk drive reliability," in *Proc. Annu. Reliab. Maintainibilty Symp.*, 1999, pp. 403–409.

[14] E. Pinheiro, W. D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population," in *Proc. Conf. FAST*, 2007, pp. 17–29.

[15] V. Cadambe, S. Jafar, H. Maleki, K. Ramchandran, and C. Suh, "Asymptotic interference alignment for optimal repair of MDS codes in distributed storage," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2974–2987, May 2013.

[16] D. J. C. MacKay, "Fountain codes," *Proc. Inst. Elect. Eng.—Commun.*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[17] L. A. Gavrilov and N. S. Gavrilova, "The reliability theory of aging and longevity," in *Handbook of the Biology of Aging*, 6th ed. San Diego, CA, USA: Academic, 2006, pp. 3–42.



Suayb S. Arslan (S'06–M'12) received the B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, San Diego, in 2009 and 2012, respectively.

He was with Mitsubishi Electric Research Laboratory, Boston, MA, in 2009, where he was involved in research and development of image and video processing algorithms for biomedical applications. In 2011, he joined Quantum Corp., Irvine, CA, where he conducted research on advanced detection and coding algorithms for increased capacity storage and cloud systems. His current research interests include digital communication and storage, joint source-channel coding, information and reliability theory, image/video processing, and cross layer design optimizations.