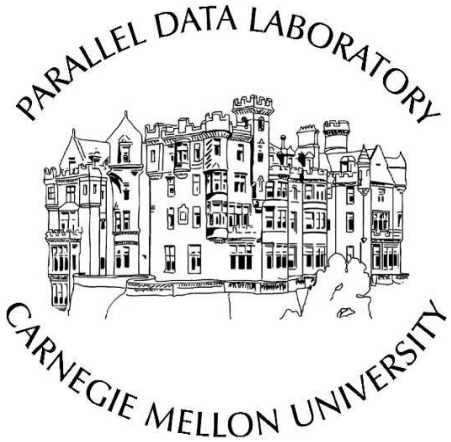


Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?

Bianca Schroeder, Garth A. Gibson

CMU-PDL-06-111
September 2006

Parallel Data Laboratory
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

Component failure in large-scale IT installations such as cluster supercomputers or internet service providers is becoming an ever larger problem as the number of processors, memory chips and disks in a single cluster approaches a million.

In this paper, we present and analyze field-gathered disk replacement data from five systems in production use at three organizations, two supercomputing sites and one internet service provider. About 70,000 disks are covered by this data, some for an entire lifetime of 5 years. All disks were high-performance enterprise disks (SCSI or FC), whose datasheet MTTF of 1,200,000 hours suggest a nominal annual failure rate of at most 0.75%.

We find that in the field, annual disk replacement rates exceed 1%, with 2-4% common and up to 12% observed on some systems. This suggests that field replacement is a fairly different process than one might predict based on datasheet MTTF, and that it can be quite variable installation to installation.

We also find evidence that failure rate is not constant with age, and that rather than a significant infant mortality effect, we see a significant early onset of wear-out degradation. That is, replacement rates in our data grew constantly with age, an effect often assumed not to set in until after 5 years of use.

In our statistical analysis of the data, we find that time between failure is not well modeled by an exponential distribution, since the empirical distribution exhibits higher levels of variability and decreasing hazard rates. We also find significant levels of correlation between failures, including autocorrelation and long-range dependence.

Acknowledgements: We thank the members and companies of the PDL Consortium (including APC, EMC, Equallogic, Hewlett-Packard, Hitachi, IBM, Intel, Microsoft, Network Appliance, Oracle, Panasas, Seagate, and Sun) for their interest, insights, feedback, and support.

Keywords: Disk failure data, failure rate, lifetime data, disk reliability, mean time to failure (MTTF), annualized failure rate (AFR).

1 Motivation

Despite major efforts, both in industry and in academia, high reliability remains a major challenge in running large-scale IT systems, and disaster prevention and cost of actual disasters make up a large fraction of the total cost of ownership. With ever larger server clusters, reliability and availability are a growing problem for many sites, including high-performance computing systems and internet service providers. A particularly big concern is the reliability of storage systems, for several reasons. First, failure of storage can not only cause temporary data unavailability, but in the worst case lead to permanent data loss. Second, many believe that technology trends and market forces may combine to make storage system failures occur more frequently in the future [19]. Finally, the size of storage systems in modern, large-scale IT installations has grown to an unprecedented scale with thousands of storage devices, making component failures the norm rather than the exception [5].

Large-scale IT systems, therefore, need better system design and management to cope with more frequent failures. One might expect increasing levels of redundancy designed for specific failure modes [2], for example. Such designs and management systems are based on very simple models of component failure and repair processes [18]. Researchers today require better knowledge about statistical properties of storage failure processes, such as the distribution of time between failures, in order to more accurately estimate the reliability of new storage system designs.

Unfortunately, many aspects of disks failures in real systems are not well understood, as it is just human nature not to advertise the details of ones failures. As a result, practitioners usually rely on vendor specified mean-time-to-failure (MTTF) values to model failure processes, although many are skeptical of the accuracy of those models [3, 4, 27]. Too much academic and corporate research is based on anecdotes and back of the envelope calculations, rather than empirical data [22].

The work in this paper is part of a broader research agenda with the long-term goal of providing a better understanding of failures in IT systems by collecting, analyzing and making publicly available a diverse set of real failure histories from large-scale production systems. In our pursuit, we have spoken to a number of large production sites and were able to convince three of them to provide failure data from several of their systems.

In this paper, we provide an analysis of five data sets we have collected, with a focus on storage-related failures. The data sets come from five different large-scale production systems at three different sites, including two large high-performance computing sites and one large internet services site. The data sets vary in duration from 1 month to 5 years and cover in total a population of more than 70,000 drives from four different vendors. All disk drives included in the data were either SCSI or fibre-channel drives, commonly represented as the most reliable types of disk drives.

We analyze the data from three different aspects. We begin in Section 3 by asking how disk failure frequencies compare to that of other hardware component failures. In Section 4, we provide a quantitative analysis of disk failure rates observed in the field and compare our observations with common predictors and models used by vendors. In Section 5, we analyze the statistical properties of disk failures. We study correlations between failures and identify the key properties of the statistical distribution of time between failures, and compare our results to common models and assumptions on disk failure characteristics.

2 Methodology

2.1 Data sources

Table 1 provides an overview over the five data sets used in this study. Data sets HPC1 and HPC2 were collected in two large cluster systems at two different organizations using supercomputers. Data sets COM1, COM2, and COM3 were collected at three different cluster systems at a large internet service provider. In

Data set	Type of cluster	Duration	Total #Events	#Disk events	# Servers	Disk Count	Disk Type	MTTF (Mhours)	System Deploym.
HPC1	HPC	08/2001 - 05/2006	1800	474	765	2,318	18GB 10K SCSI	1.2	08/2001
			463	124	64	1,088	36GB 10K SCSI	1.2	
HPC2	HPC	01/2004 - 07/2006	14	14	256	520	36GB 10K SCSI	1.2	12/2001
COM1	Int. serv.	May 2006	465	84	N/A	26,734	10K SCSI	1	2001
COM2	Int. serv.	09/2004 - 04/2006	667	506	9,232	39,039	15K SCSI	1.2	2004
COM3	Int. serv.	01/2005 - 12/2005	104	104	N/A	432	10K FC-AL	1.2	1998
			2	2	N/A	56	10K FC-AL	1.2	N/A
			132	132	N/A	2,450	10K FC-AL	1.2	N/A
			108	108	N/A	796	10K FC-AL	1.2	N/A

Table 1: *Overview of the five failure data sets*

all cases, our data reports on only a portion of the computing systems run by each organization. Below we describe each data set and the system it comes from in some more detail.

HPC1 is a five year log of hardware failures collected from a 765 node high-performance computing cluster. Each of the 765 nodes is a 4-way SMP with 4 GB of memory and 3-4 18GB 10K rpm SCSI drives. 64 of the nodes are used as filesystem nodes containing in addition to the 3-4 18GB drives, 17 36GB 10K rpm SCSI drives. The applications running on those systems are typically large-scale scientific simulations or visualization applications. The data contains, for each hardware failure that was recorded during the 5 year lifetime of this system, when the problem started, which node and which hardware component was affected, and a brief description of the corrective action.

HPC2 is a record of disk failures observed on the compute nodes of a 256 node HPC cluster. Each node is a 4-way SMP with 16 GB of memory and contains two 36GB 10K rpm SCSI drives, except for 8 of the nodes, which contain eight 36GB 10K rpm SCSI drives each. The applications running on those systems are typically large-scale scientific simulations or visualization applications. For each disk failure the data set records the number of the affected node, the start time of the failure, and the slot number of the failed drive.

COM1 is a log of hardware failures recorded at a cluster at an internet service provider. Each failure record in the data contains a timestamp on when the failure was repaired, information on the failure symptoms, and a list of steps that were taken to repair the problem. Note that this data does not contain information on when a failure actually happened, only when repair took place. The data covers a population of 26,734 10K SCSI disk drives. The number of servers in the environment is not known.

COM2 is also a vendor-created log of hardware failures recorded at a cluster at an internet service provider. Each failure record contains a repair code (e.g. "Replace hard drive") and the time when the repair was finished. Again there's no information on the start time of a failure. The log does not contain entries for failures of disks that were replaced as hot-swaps, since the data was created by the vendor, who doesn't see those replacements. To account for the missing disk replacements we obtained numbers for disk replacements from the internet service provider. The size of the underlying system changed significantly during the measurement period, starting with 420 servers in 2004 and ending with 9,232 servers in 2006. We obtained hardware purchase records for the system for this time period to estimate the size of the disk population for each quarter of the measurement period.

The COM3 dataset comes from a large storage system at an internet service provider and comprises four populations of different types of fibre-channel disks (see Table 1). While this data was gathered in 2005, the system has some legacy components that are as old as from 1998. COM3 differs from the other data sets in that it provides only aggregate statistics of disk failures, rather than individual records for each failure. The data contains the counts of disks that failed and were replaced in 2005 for each of the four disk populations.

2.2 Statistical methods

We characterize an empirical distribution using two import metrics: the mean, and the squared coefficient of variation (C^2). The squared coefficient of variation is a measure of the variability of a distribution and is defined as the squared standard deviation divided by the squared mean. The advantage of using the squared coefficient of variation as a measure of variability, rather than the variance or the standard deviation, is that it is normalized by the mean, and hence allows comparison of variability across distributions with different means.

We also consider the empirical cumulative distribution function (CDF) and how well it is fit by four probability distributions commonly used in reliability theory[1]: the exponential distribution; the Weibull distribution; the gamma distribution; and the lognormal distribution. We parameterize the distributions through maximum likelihood estimation and evaluate the goodness of fit by visual inspection, the negative log-likelihood and the chi-square test.

Since we are interested in correlations between disk failures we need a measure for the degree of correlation. The autocorrelation function (ACF) measures the correlation of a random variable with itself at different time lags l. The ACF can for example be used to determine whether the number of failures in one day is correlated with number of failures observed l days later. The autocorrelation coefficient can range between 1 (high positive correlation) and -1 (high negative correlation).

Another aspect of the failure process that we will study is long-range dependence. Long-range dependence measures the memory of a process, in particular how quickly the autocorrelation coefficient decays with growing lags. The strength of the long-range dependence is quantified by the Hurst exponent. A series exhibits long-range dependence if the Hurst exponent H is $0.5 < H < 1$. We use the Selfis tool [12] to obtain estimates of the Hurst parameter using five different methods: the absolute value method, the variance method, the R/S method, the periodogram method, and the Whittle estimator.

3 Comparing failures of disks to other hardware components

The reliability of a system depends on all its components, and not just the hard drive(s). A natural question is therefore what the relative frequency of drive failures is, compared to that of other types of hardware failures. To answer this question we consult data sets HPC1, COM1, and COM3, since these data sets contain records for any type of hardware failure, not only disk failures.

We begin by considering only *permanent* hardware failures, i.e. failures that require replacement of the affected hardware component. Table 2 shows, for each data set, a list of the ten most frequently replaced hardware components and the fraction of replacements made up by each component. We observe that while the actual fraction of disk replacements varies across the data sets (ranging from 20% to 50%), it makes up a significant fraction in all three cases. In the HPC1 and COM2 data sets, disk drives are the most commonly replaced hardware component accounting for 30% and 50% of all hardware replacements, respectively. In the COM1 data set, disks are a close runner-up accounting for nearly 20% of all hardware replacements.

While Table 2 suggests that disks are among the most commonly replaced hardware components, it does not necessarily imply that disks are less reliable or have a shorter lifespan than other hardware components. The number of disks in the systems might simply be much larger than that of other hardware components. In order to compare the reliability of different hardware components, we need to normalize the number of component replacements by the component's populations size.

Unfortunately, we do not have, for any of the five systems, exact population counts of all hardware components. However, we do have enough information on the HPC1 system to estimate counts of the four

[1]We also considered another distribution, which has recently been found to be useful in characterizing various aspects of computer systems, the Pareto distribution. However, we didn't find it to be a better fit than any of the four standard distributions for our data and therefore did not include it in these results.

HPC1	
Component	%
Hard drive	**30.6**
Memory	28.5
Misc/Unk	14.4
CPU	12.4
PCI motherboard	4.9
Controller	2.9
QSW	1.7
Power supply	1.6
MLB	1.0
SCSI BP	0.3

COM1	
Component	%
Power supply	34.8
Memory	20.1
Hard drive	**18.1**
Case	11.4
Fan	8.0
CPU	2.0
SCSI Board	0.6
NIC Card	1.2
LV Power Board	0.6
CPU heatsink	0.6

COM2	
Component	%
Hard drive	**49.1**
Motherboard	23.4
Power supply	10.1
RAID card	4.1
Memory	3.4
SCSI cable	2.2
Fan	2.2
CPU	2.2
CD-ROM	0.6
Raid Controller	0.6

Table 2: *Relative frequency of hardware component failures that require replacement*

hardware components that fail most frequently in this system (CPU, memory, disks, motherboards). We estimate that there is a total of 3,060 CPUs, 3,060 dimms, and 765 mother boards, compared to a disk population of 3,406. Combining these numbers with the data in Table 2, we conclude that for the HPC1 system, the probability that in 5 years of use a memory dimm will be replaced is roughly comparable to that of a hard drive replacement; a CPU is about 2.5 times less likely to be replaced than a hard drive; and a motherboard is 50% less likely to be replaced than a hard drive.

The above discussion focused only on permanent failures, i.e. failures that required a hardware component to be replaced. When running a large system one is also interested in other hardware failures, that cause a node to go down. The HPC1 data also contains records for hardware related failures that caused a node outage, but did not require hardware replacement. Table 3 gives a per-component breakdown of all hardware failures in HPC1, including those that did not require hardware replacement. We observe that the percentage of failures caused by disk drives is 16% (compared to 30% in Table 2), making it the third most common hardware-related root cause of a node outage. Memory failures are nearly two times more common than disk failures, and CPU failures are almost three times more common than disk failures.

HPC1	
Component	%
CPU	44
Memory	29
Hard drive	**16**
PCI motherboard	9
Power supply	2

Table 3: *Relative frequency of hardware-related failures, including those that did not require component replacement*

For a complete picture, we also need to take the severity of a failure into account. A closer look at the HPC1 data reveals that a large number of the CPU and memory failures are triggered by parity errors, i.e. the number of errors is too large for the embedded error correcting code to correct them. Those errors just require a reboot to bring the affected node back up. On the other hand, the majority of the disk failures (around 90%) recorded in HPC1 was permanent, requiring more time-consuming and expensive repair actions.

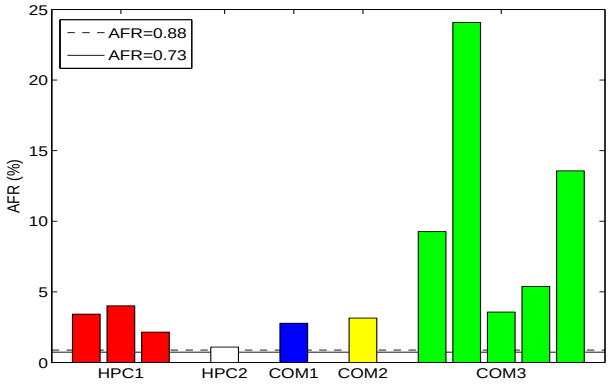


Figure 1: *Comparison of datasheet AFRs (solid and dashed line in the graph) and AFRs observed in the field (bars in the graph). Left-most bar in a set is the result of combining all types of disks in the data set.*

4 Disk failure rates

4.1 Specifying disk reliability and failure frequency

Drive manufacturers specify the reliability of their products in terms of two related metrics: the *annualized failure rate (AFR)*, which is the percentage of disk drives in a population that fail in a test scaled to a per year estimation; and the *mean time to failure (MTTF)*. The AFR of a new product is typically estimated based on accelerated life and stress tests or based on field data from earlier products [1]. The MTTF is estimated as the number of power on hours per year[2] divided by the AFR. The MTTFs specified for today's highest quality disks range from 1,000,000 hours to 1,400,000 hours, corresponding to AFRs of 0.63% to 0.88%.

The AFR and MTTF estimates of the manufacturer are included in a drive's datasheet and we refer to them in the remainder as the *datasheet AFR* and the *datasheet MTTF*. In contrast, we will refer to the AFR and MTTF computed from the data sets as the *observed AFR* and *observed MTTF*, respectively.

4.2 Disk failures and MTTF

In the following, we study how field experience with disk failures compares to datasheet specifications of disk reliability. Figure 1 shows the datasheet AFRs (horizontal solid and dashed line) and the observed AFR for each of the five data sets. For HPC1 and COM3, which cover different types of disks, the graph contains several bars, one for the observed AFR across all types of disk (left-most bar), and one for the AFR of each type of disk (remaining bars in the order of the corresponding entries in Table 1).

We observe a significant discrepancy between the observed AFR and the datasheet AFR for all data sets. While the datasheet AFRs are either 0.73% or 0.88%, the observed AFRs range from from 1.1% to as high as 25%. That is the observed AFRs are by a factor of 1.5 to up to a factor of 30 higher than the datasheet AFRs.

A striking observation in Figure 1 is the huge variation of AFRs across the systems, in particular the extremely large AFRs observed in system COM3. While 3,302 of the disks in COM3 were at all times less than 5 years old, 432 of the disks in this system were installed in 1998, making them at least 7 years old at the end of the data set. Since this is well outside the vendor's nominal lifetime for disks, it is not surprising that the disks might be wearing out. But even without the 432 obsolete disks, COM3 has quite large AFRs.

[2]A common assumption for enterprise drives is that they are 100% of the time powered on. Our data set providers all believe that their disks are 100% powered on.

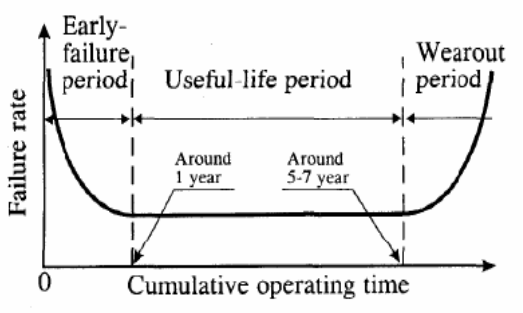


Figure 2: *Life cycle failure pattern of hard drives [27].*

The data for HPC1 covers almost exactly 5 years, the nominal lifetime, and exhibits an AFR significantly higher than the datasheet AFR (3.4% instead of 0.88%). The data for COM2 covers the first 2 years of operation and has an AFR of 3.1%, also much higher than the datasheet AFR of 0.88%.

It is interesting to observe that the only system that comes close to the datasheet AFR is HPC2, which with an observed AFR of 1.1%, deviates from the datasheet AFR by "only" 50%. After talking to people involved in running system HPC2, we identified as a possible explanation the potentially very low usage of the disks in HPC2. The disks in this data set are local disks on compute nodes, whose applications primarily use a separate, shared parallel file system, whose disks are not included in the data set. The local disks, which are included in the data set, are mostly used only for booting to the operating system, and fetching system executables/libs. Users are allowed to write only to a smallish /tmp area of the disks and are thought to do this rarely, and swapping almost never happens.

Below we summarize the key observations of this section.

Observation 1: Variance between datasheet MTTF and field failure data is larger than one might expect.

Observation 2: For older systems (5-8 years of age), data sheet MTTFs can underestimate failure rates by as much as a factor of 30.

Observation 3: Even during the first few years of a system's lifetime (< 3 years), when wear-out is not expected to be a significant factor, the difference between datasheet MTTF and observed MTTF can be as large as a factor of 6.

4.3 Age-dependent failure rates

One aspect of failure rates that single-value metrics such as MTTF and AFR cannot capture is that in real life failure rates are not constant [4]. Failure rates of hardware products typically follow a "bathtub curve" with high failure rates at the beginning (infant mortality) and the end (wear-out) of the lifecycle. Figure 2 shows the failure rate pattern that is expected for the life cycle of hard drives [3, 4, 27]. According to this model, the first year of operation is characterized by early failures (or infant mortality). In years 2-5, the failure rates are approximately in steady state, and then, after years 5-7, wear-out starts to kick in.

The common concern, that MTTFs don't capture infant mortality, has lead the International Disk drive Equipment and Materials Association (IDEMA) to propose a new standard for specifying disk drive reliability, based on the failure model depicted in Figure 2 [4]. The new standard requests that vendors provide four different MTTF estimates, one for the first 1-3 months of operation, one for months 4-6, one for months

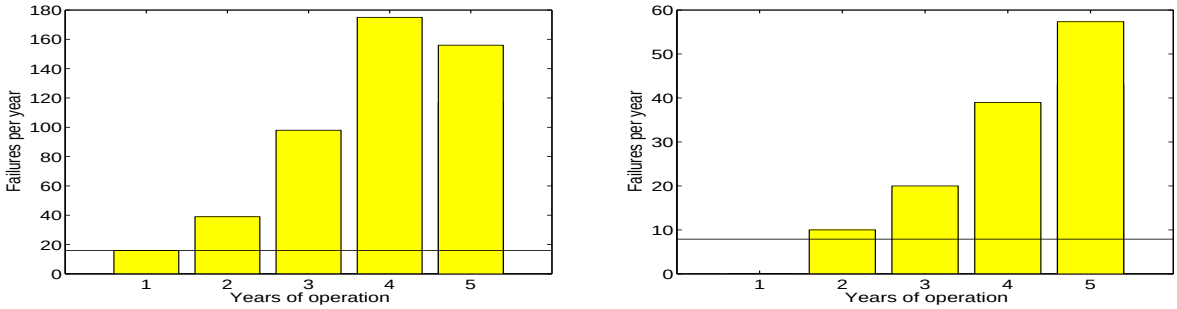


Figure 3: *Number of failures observed per year over the first 5 years of system HPC1's lifetime, for the compute nodes (left) and the file system nodes (right).*

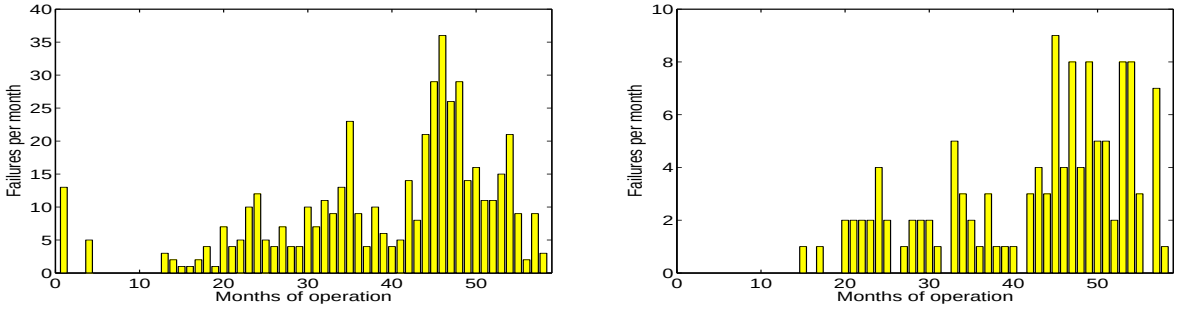


Figure 4: *Number of failures observed per month over the first 5 years of system HPC1's lifetime, for the compute nodes (left) and the file system nodes (right).*

7-12, and one for months 13-60.

The goal of this section is to study, based on our field replacement data, how failure rates in large-scale installations vary over a system's life cycle.

The best data set to study failure rates across the system life cycle is system HPC1. The reason is that this data set spans the entire first 5 years of operation of a large system. Moreover, in HPC1 the hard drive population is homogeneous, with all 3,406 drives in the system being nearly identical (except for having two different sizes, 17 vs 36 GB), and the population size remained the same over the 5 years, except for the small fraction of replaced components.

We study the change of failure rates across system HPC1's lifecycle at two different granularities, on a per-month and a per-year basis, to make it easier to detect both short term and long term trends. Figure 3 shows the yearly failure rates for the disks in the compute nodes of system HPC1 (left) and the file system nodes of system HPC1 (right). We make two interesting observations. First, failure rates in all years, except for year 1, are dramatically larger than the datasheet MTTF would suggest. The solid line in the graph represents the number of failures expected per year based on the data sheet MTTF. In year 2, disk failure rates are 20% larger than expected for the file system nodes, and a factor of two larger than expected for the compute nodes. In year 4 and year 5 (which are still within the nominal lifetime of these disks), the actual failure rates are 7–10 times higher than expected.

The second observation is that failure rates are rising significantly over the years, even during early years in the lifecycle. Failure rates nearly double when moving from year 2 to 3 or from year 3 to 4. This observation suggests that wear-out may start much earlier than expected, leading to steadily increasing failure rates during most of a system's useful life. This is an interesting observation because it does not agree with the common assumption that after the first year of operation, failure rates reach a steady state for

a few years, forming the "bottom of the bathtub".

Next, we move to the per-month view of system HPC1's failure rates, shown in Figure 4. We observe that for the file system nodes, there's is no detectable infant mortality: there are no failures observed during the first 12 months of operation. In the case of the compute nodes, infant mortality is limited to the first month of operation and is not above the steady state estimate of the datasheet MTTF. Looking at the life-cycle after month 12, we again see continuously rising failure rates, instead of the expected "bottom of the bathtub".

Below we summarize the key observations of this section.

Observation 4: Contrary to common and proposed models, hard drive failure rates don't enter steady state after the first year of operation. Instead failure rates seem to steadily increase over time.

Observation 5: Early onset of wear-out seems to have a much stronger impact on lifecycle failure rates than infant mortality, even when considering only the first 3 or 5 years of a system's lifetime. Wear-out should therefore be a incorporated into new standards for disk drive reliability. The new standard suggested by IDEMA does not take wear-out into account.

5 Statistical properties of disk failures

In the previous sections, we have focused on aggregate failure statistics, e.g. the average failure rate in a time period. Often one wants more information on the statistical properties of the time between failures than just the mean. For example, determining the expected time to failure for a RAID system requires an estimate on the probability of experiencing a second disk failure in a short period, that is while reconstructing lost data from redundant data. This probability depends on the underlying probability distribution and maybe poorly estimated by scaling an annual failure rate down to a few hours.

The most common assumption about the statistical characteristics of disk failures is that they form a Poisson process, which implies two key properties:

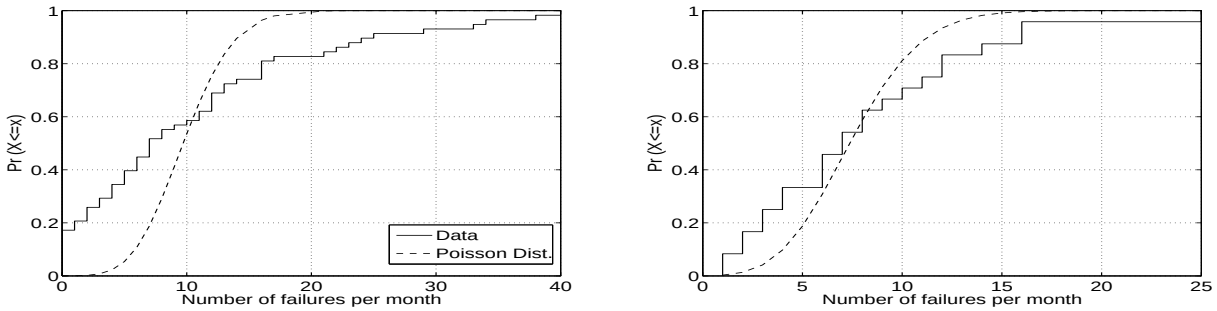
1. Failures are independent.
2. The time between failures follows an exponential distribution.

The goal of this section is to evaluate how realistic the above assumptions are. We begin by providing statistical evidence that disk failures in the real world are unlikely to follow a Poisson process. We then examine in Section 5.2 and Section 5.3 each of the two key properties (independent failures and exponential time between failures) independently and characterize in detail how and where the Poisson assumption breaks. In our study, we focus on the HPC1 data set, since this is the only data set that contains precise failure time stamps (rather than just repair time stamps).

5.1 The Poisson assumption

The Poisson assumption implies that the number of failures during a given time interval (e.g. a week or a month) is distributed according to the Poisson distribution. Figure 5 (left) shows the empirical CDF of the number of failures observed per month in the HPC1 data set, together with the Poisson distribution fit to the data's observed mean.

We find that the Poisson distribution does not provide a good fit for the number of failures observed per month in the data, in particular for very small and very large numbers of failures. For example, under the Poisson distribution the probability of seeing ≥ 20 failures in a given month is less than 0.0024, yet we



All years Years 2-3

Figure 5: *CDF of number of failures per month in HPC1*

see 20 or more failures in nearly 20% of all months in HPC1's lifetime. Similarly, the probability of seeing zero or one failure in a given month is only 0.0003 under the Poisson distribution, yet in 20% of all months in HPC1's lifetime we observe zero or one failure.

A chi-square test reveals that we can reject the hypothesis that the number of failures per month follows a Poisson distribution at the 0.05 significance level. All above results are similar when looking at the distribution of failures per day or per week, rather than per month.

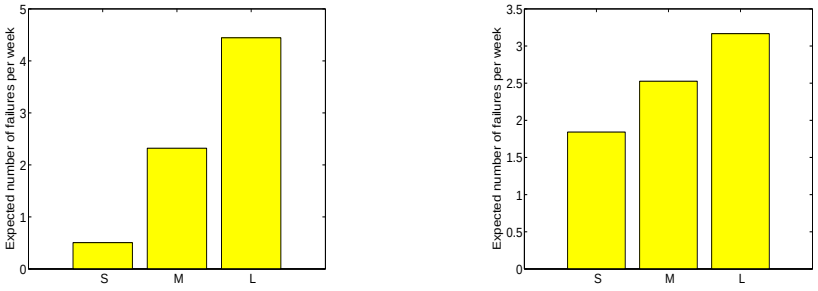
One reason for the poor fit of the Poisson distribution might be that failure rates are not steady over the lifetime of HPC1. We therefore repeat the same process for only part of HPC1's lifetime. Figure 5 (right) shows the distribution of failures per month, using only data from years 2 and 3 of HPC1. The Poisson distribution achieves a better fit for this time period and the chi-square test cannot reject the Poisson hypothesis at a significance level of 0.05. Note, however, that this does not necessarily mean that the failure process during years 2 and 3 does follow a Poisson process, since this would also require the two key properties of a Poisson process (independent failures and exponential time between failures) to hold. We study these two properties in detail in the next two sections.

5.2 Correlations

In this section, we focus on the first key property of a Poisson process, the independence of failures. Intuitively, it is clear that in practice failures of disks in the same system are never completely independent. The failure probability of disks depends for example on environmental factors, such as temperature, that are shared by all disks in the system. When the temperature in a machine room rises, all disks in the room experience a higher than normal probability of failure. The goal of this section is to statistically quantify and characterize the correlation between disk failures.

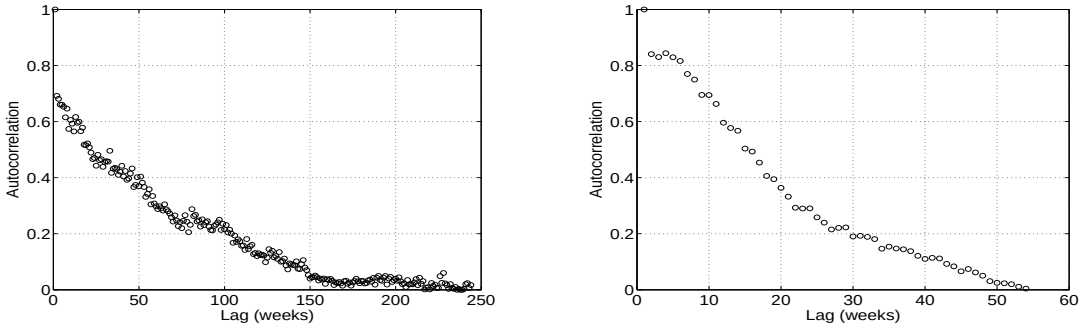
We start with a simple test in which we determine the correlation of the number of failures observed in successive weeks/months by computing the correlation coefficient between the number of failures in a given week/month and the previous week/month. For data coming from a Poisson processes we would expect correlation coefficients close to 0. Instead we find significant levels of correlations, both at the month and the week level. The correlation coefficient between consecutive weeks is 0.72 the correlation coefficient between consecutive months is 0.79. We repeated the same test using only the data of one year at a time, and we still find significant levels of correlation with correlation coefficients of 0.4-0.8.

Statistically, the above correlation coefficients indicate a strong correlation, but it would be nice to have a more intuitive interpretation of this result. One way of thinking of the correlation of failure rates is that the failure rate in one time interval is predictive of the failure rate in the following time interval. To test the strength of this prediction, we assign each week in HPC1's life to one of three buckets, depending on the



All years Year 3

Figure 6: *Expected number of failures in a week depending depending on the number of failures in the previous week.*



All years Year 3

Figure 7: *Autocorrelation function for the number of failures per week computed across the entire lifetime of the HPC1 system (left) and computed across only one year of HPC1's operation (right).*

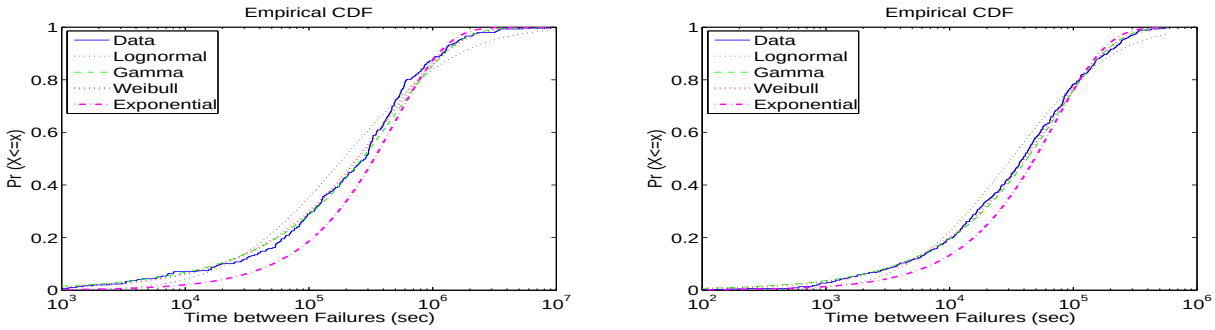
number of failures observed during that week, creating a bucket for weeks with small, medium, and large number of failures, respectively [3]. The expectation is that a week that follows a week with a "small" number of failures is more likely to see a small number of failures, than a week that follows a week with a "large" number of failures. Figure 6 (left) shows the expected number of failures in a week of HPC1's lifetime as a function of which bucket the preceding week falls in. We observe that the expected number of failures in a week varies by a factor of 9, depending on whether the preceding week falls into the first or third bucket. When repeating the same process on the data of only year 3 of HPC1's lifetime, we see a differences of a close to factor of 2 between the first and third bucket.

So far, we have only considered correlations between successive time intervals, e.g. between two successive weeks. A more general way to characterize correlations is to study correlations at different time lags by using the autocorrelation function. Figure 7 (left) shows the autocorrelation function for the number of failures per week computed across the HPC1 data set. For a stationary failure process (e.g. data coming from a Poisson process) the autocorrelation would be close to zero at all lags. Instead, we observe strong autocorrelation even for large lags in the range of 100 weeks (nearly 2 years).

We also repeated the same test for only parts of HPC1's lifetime and find similar levels of autocorrelation. Figure 7 (right), for example, shows the autocorrelation function computed only on the data of the third year of HPC1's life. Correlation is significant for lags in the range of up to 30 weeks.

Another measure for dependency is long range dependence, as quantified by the Hurst exponent H.

[3] More precisely, we choose the cutoffs between the buckets such that each bucket contains the same number of samples (i.e. weeks) by using the 33th percentile and the 66th percentile of the empirical distribution as cutoffs between the buckets.



All years Year 3

Figure 8: *Distribution of time between failures across all nodes in HPC1.*

The Hurst exponent measures how fast the autocorrelation functions drops with increasing lags. A Hurst parameter between 0.5–1 signifies a statistical process with a long memory and a slow drop of the autocorrelation function. Applying several different estimators (see Section 2) to the HPC1 data, we determine a Hurst exponent between 0.6-0.8 at the weekly granularity. These values are comparable to Hurst exponents reported for Ethernet traffic, which is known to exhibit strong long range dependence.

Observation 6: Disk failures exhibit significant levels of autocorrelation.

Observation 7: Disk failures exhibit long-range dependence.

5.3 Distribution of time between failure

In this section, we focus on the second key property of a Poisson failure process, the exponentially distributed time between failures. Figure 8 (left) shows the empirical cumulative distribution function of time between failures as observed in the HPC1 system and four distributions matched to it.

We find that visually the Gamma and Weibull distributions are the best fit to the data, while exponential and lognormal distributions provide a poorer fit. This agrees with results we obtain from the negative log-likelihood, which indicates that the Weibull distribution is the best fit, closely followed by the gamma distribution. Performing a Chi-Square-Test, we can reject the hypothesis that the underlying distribution is exponential or lognormal at a significance level of 0.05. On the other hand the hypothesis that the underlying distribution is a Weibull or a gamma cannot be rejected at a significance level of 0.05.

The poor fit of the exponential distribution might be due to the fact that failure rates change over the lifetime of the system, creating variability in the observed times between failure that the exponential distribution cannot capture. We therefore repeated the above analysis considering only segments of HPC1's lifetime. Figure 8 (right) shows as one example the results from analyzing the time between failures in year 3 of HPC1's operation. While visually the exponential distribution now seems a slightly better fit, we can still reject the hypothesis of an underlying exponential distribution with a significance level of 0.05. The same holds for other 1-year and even 6-month segments of HPC1's lifetime. This leads us to conclude that even during shorter segments of HPC1's lifetime the time between failures is not realistically modeled by an exponential distribution.

While it might not come as a surprise that the simple exponential distribution does not provide as good a fit as the more flexible two-parameter distributions, an interesting question is what properties of the empirical time between failure make it different from a theoretical exponential distribution. We identify as a first differentiating feature that the data exhibits higher variability than a theoretical exponential distribution.

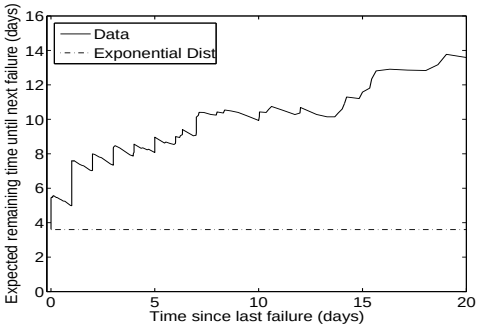


Figure 9: *Illustration of decreasing hazard rates*

The data has a C^2 of 2.4, which is more than two times higher than the C^2 of an exponential distribution which is 1.

A second differentiating feature is that the time between failure in the data exhibits decreasing hazard rates, as indicated by the shape parameters of the fitted Gamma and Weibull distributions (shape parameter less than 1). The hazard rate measures how the time since the last failure influences the expected time until the next failure. An increasing hazard rate function predicts that if the time since a failure is long then the next failure is coming soon. And a decreasing hazard rate function predicts the reverse.

Figure 9 illustrates the data's decreasing hazard rates by plotting the expected remaining time until the next failure (Y-axis) as a function of the time since the last failure (X-axis). We observe that right after a failure the expected time until the next failure is around 4 days, both for the empirical data and the exponential distribution. In the case of the empirical data, after surviving for 10 ten days without failures the expected remaining time until the next failure grows from initially 4 to 10; and after surviving for a total of 20 days without failures the expected time until the next failure grows to 15 days. In comparison, under an exponential distribution the expected remaining time stays constant (also known as the memoryless property).

Observation 8: The hypothesis that time between failures follows an exponential distribution can be rejected with high confidence.

Observation 9: The time between failures has a higher variability than that of an exponential distribution.

Observation 10: The distribution of time between failures exhibits decreasing hazard rates, that is the expected remaining time until the next failures grows with the time since we have seen the last failure.

6 Related work

There is very little work published on analyzing failures in real, large-scale storage systems, probably as a result of the reluctance of the owners of such systems to gather and release failure data.

Among the few existing studies is the work by Talagala et al. [23], which provides a study of error logs in a research prototype storage system used for a web server and includes a comparison of failure rates of different hardware components. They identify SCSI disk enclosures as the least reliable components and SCSI disks as one of the most reliable component, which differs from our results.

In a recently initiated effort, Schwarz et al. [22] have started to gather failure data at the Internet Archive, which they plan to use to study disk failure rates and bit rot rates and how they are affected by

different environmental parameters. In their preliminary results, they report AFR values of 2–6% and note that the Archive does not seem to see significant infant mortality. Both observations are in agreement with our findings.

Gray [25] reports the frequency of uncorrectable read errors in disks and finds that their numbers are smaller than vendor data sheets suggest. Gray also provides AFR estimates for SCSI and ATA disks, in the range of 3–6%, which is in the range of AFRs that we observe for SCSI drives in our data sets.

Many have criticized the accuracy of MTTF based failure rate predictions and have pointed out the need for more realistic models. A particular concern is the fact that a single MTTF value cannot capture life cycle patterns [3, 4, 27]. Our analysis of life cycle patterns shows that this concern is justified, since we find failure rates to vary quite significantly over even the first 2-3 years of the life cycle. However, the most common life cycle concern in published research is underrepresenting infant mortality. In our analysis, we don't see that. Instead we observe significant underrepresentation of the early onset of wear-out.

Early work on RAID systems [6] provided some statistical analysis of time between disk failures for disks used in the 1980s, but didn't find sufficient evidence to reject the hypothesis of exponential times between failure with high confidence. However, time between failure has been analyzed for other, non-storage data in several studies [9, 13, 20, 21, 24, 26]. Four of the studies use distribution fitting and find the Weibull distribution to be a good fit [9, 13, 21, 26], which agrees with our results. All studies looked at the hazard rate function, but come to different conclusions. Four of them [9, 13, 21, 26] find decreasing hazard rates (Weibull shape parameter < 0.5). Others find that hazard rates are flat [24], or increasing [20]. We find decreasing hazard rates with Weibull shape parameter of 0.7-0.8. Disk vendors, in fact, use a Weibull model to derive the datasheet MTTF based on accelerated/stress testing over short periods of time [1].

Large-scale failure studies are scarce, even when considering IT systems in general and not just storage systems. Most existing studies are limited to only a few months of data, covering typically only a few hundred failures [11, 16, 17, 20, 24, 26]. And many of the most commonly cited studies on failure analysis stem from the late 80's and early 90's, when computer systems where significantly different from today [7, 8, 10, 13, 14, 15, 24].

7 Conclusion

Many have pointed out the need for a better understanding of what disk failures look like in the field. Yet hardly any published work exists that provides large-scale studies of disk failures in production systems. As a first step towards closing this gap, we have analyzed disk failure data from five different large production systems, spanning more than 70,000 drives from four different vendors, including both SCSI and fibre-channel drives. Below is a summary of a few of our results.

- Large-scale installation field usage appears to differ widely from nominal datasheet MTTF conditions. The field failure rates of systems are significantly larger than one would expect based on datasheet MTTFs.

- For less than 5 year old drives, field failure rates are by a factor of 2–12 larger than what the datasheet MTTF suggests. For 5-8 year old drives, field failure rates can be as much as a factor of 30 higher than what the datasheet MTTF suggests.

- Changes in failure rates during the first 5 years of the lifecycle are more dramatic than often assumed. While failure rates are often expected to be in steady state in year 2-5 of operation (bottom of the "bathtub curve"), we observe a continuous increase in failure rates, starting as early as in the second year of operation.

- The common concern that MTTFs underrepresent infant mortality has led to the proposal of new standards that incorporate infant mortality [27]. We find that the underrepresentation of the early onset of wear-out is a much more serious factor than underrepresentation of infant mortality and recommend to include this in new standards.
- While many have suspected that the commonly made assumption of exponentially distributed time between failures is not realistic, previous studies have not found enough evidence to prove this assumption wrong with significant statistical confidence [6]. Based on our data analysis, we are able to reject the hypothesis of exponentially distributed time between failures with high confidence.
- We identify as the key features that distinguish the empirical time between failure distribution from the exponential distribution, a higher levels of variability and decreasing hazard rates. We find that the empirical distributions are fit well by a Weibull distribution with shape parameter less than 1.
- We also present strong evidence for the existence of correlations between failures. In particular, the empirical data exhibits significant levels of autocorrelation and long-range dependence.

References

[1] G. Cole. Estimating drive reliability in desktop computers and consumer electronics systems. TP-338.1. Seagate. 2000.

[2] Peter F. Corbett, Robert English, Atul Goel, Tomislav Grcanac, Steven Kleiman, James Leong, and Sunitha Sankar. Row-diagonal parity for double disk failure correction. In *Proceedings of the FAST '04 Conference on File and Storage Technologies*, 2004.

[3] John G. Elerath. AFR: problems of definition, calculation and measurement in a commercial environment. In *2000 Proceedings Annual Reliability and Maintainability Symposium*, 2000.

[4] John G. Elerath. Specifying reliability in the disk drive industry: No more MTBFs. In *2000 Proceedings Annual Reliability and Maintainability Symposium*, 2000.

[5] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 29–43. ACM Press, 2003.

[6] Garth A. Gibson. Redundant disk arrays: Reliable, parallel secondary storage. Dissertation. MIT Press. 1992.

[7] J. Gray. Why do computers stop and what can be done about it. In *Proc. of the 5th Symposium on Reliability in Distributed Software and Database Systems*, 1986.

[8] J. Gray. A census of tandem system availability between 1985 and 1990. *IEEE Transactions on Reliability*, 39(4), 1990.

[9] T. Heath, R. P. Martin, and T. D. Nguyen. Improving cluster availability using workstation validation. In *Proc. of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, 2002.

[10] R. K. Iyer, D. J. Rossetti, and M. C. Hsueh. Measurement and modeling of computer reliability as affected by system activity. *ACM Trans. Comput. Syst.*, 4(3), 1986.

[11] M. Kalyanakrishnam, Z. Kalbarczyk, and R. Iyer. Failure data analysis of a LAN of Windows NT based computers. In *Proc. of the 18th IEEE Symposium on Reliable Distributed Systems*, 1999.

[12] T. Karagiannis. Selfis: A short tutorial. Technical report, University of California, Riverside, 2002.

[13] T.-T. Y. Lin and D. P. Siewiorek. Error log analysis: Statistical modeling and heuristic trend analysis. *IEEE Transactions on Reliability*, 39, 1990.

[14] J. Meyer and L. Wei. Analysis of workload influence on dependability. In *Proc. International Symposium on Fault-tolerant computing*, 1988.

[15] B. Murphy and T. Gent. Measuring system and software reliability using an automated data collection process. *Quality and Reliability Engineering International*, 11(5), 1995.

[16] D. Nurmi, J. Brevik, and R. Wolski. Modeling machine availability in enterprise and wide-area distributed computing environments. In *Euro-Par'05*, 2005.

[17] D. L. Oppenheimer, A. Ganapathi, and D. A. Patterson. Why do internet services fail, and what can be done about it? In *USENIX Symposium on Internet Technologies and Systems*, 2003.

[18] David Patterson, Garth Gibson, and Randy Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proc. of the ACM SIGMOD International Conference on Management of Data*, 1988.

[19] Vijayan Prabhakaran, Lakshmi N. Bairavasundaram, Nitin Agrawal, Haryadi S. Gunawi, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Iron file systems. In *SOSP '05: Proceedings of the twentieth ACM symposium on Operating systems principles*, pages 206–220, 2005.

[20] R. K. Sahoo, R. K., A. Sivasubramaniam, M. S. Squillante, and Y. Zhang. Failure data analysis of a large-scale heterogeneous server environment. In *Proc. of the 2004 international Conference on Dependable Systems and Networks (DSN'04)*, 2004.

[21] B. Schroeder and G. Gibson. A large-scale study of failures in high-performance computing systems. In *Proc. of the 2006 international Conference on Dependable Systems and Networks (DSN'06)*, 2006.

[22] T. Schwarz, M. Baker, S. Bassi, B. Baumgart, W. Flagg, C. van Ingen, K. Joste, M. Manasse, and M. Shah. Disk failure investigations at the internet archive. In *Work-in-Progess session, NASA/IEEE Conference on Mass Storage Systems and Technologies (MSST2006)*, 2006.

[23] Nisha Talagala and David Patterson. An analysis of error behaviour in a large storage system. In *The IEEE Workshop on Fault Tolerance in Parallel and Distributed Systems*, 1999.

[24] D. Tang, R. K. Iyer, and S. S. Subramani. Failure analysis and modelling of a VAX cluster system. In *Proc. International Symposium on Fault-tolerant computing*, 1990.

[25] C. van Ingen and J. Gray. Empirical measurements of disk failure rates and error rates. In *MSR-TR-2005-166*, 2005.

[26] J. Xu, Z. Kalbarczyk, and R. K. Iyer. Networked Windows NT system field failure data analysis. In *Proc. of the 1999 Pacific Rim International Symposium on Dependable Computing*, 1999.

[27] Jimmy Yang and Feng-Bin Sun. A comprehensive review of hard-disk drive reliability. In *1999 Proceedings Annual Reliability and Maintainability Symposium*, 1999.