# Reliability of MEMS-Based Storage Enclosures

Bo Hong[1†] *hongbo@cs.ucsc.edu*
Thomas J. E. Schwarz, S. J.[2‡] *tjschwarz@scu.edu*
Scott A. Brandt[1†] *scott@cs.ucsc.edu*
Darrell D. E. Long[1†] *darrell@cs.ucsc.edu*

[1]Storage Systems Research Center, University of California, Santa Cruz, CA 95064
[2]Computer Engineering Department, Santa Clara University, Santa Clara, CA 95053

## Abstract

*MEMS-based storage is a new, non-volatile storage technology currently under development. It promises fast data access, high throughput, high storage density, small physical size, low power consumption, and low entry costs. These properties make MEMS-based storage into a serious alternative to disk drives, in particular for mobile applications. The first generation of MEMS will only offer a fraction of the storage capacity of disks; therefore, we propose to integrate multiple MEMS devices into a* MEMS storage enclosure, *organizing them as a RAID Level 5 with multiple spares, to be used as the basic storage building block. This paper investigates the reliability of such an enclosure. We find that Mean Time To Failure is an inappropriate reliability metric for MEMS enclosures. We show that the reliability of the enclosures is appropriate for their economic lifetime if users choose not to replace failed MEMS storage components. In addition, we investigate the benefits of occasional, preventive maintenance of enclosures.*

## 1. Introduction

Magnetic disks have dominated secondary storage for decades. A new class of secondary storage devices based on microelectromechanical systems (MEMS) is a promising non-volatile secondary storage technology currently being developed [4, 18, 27, 28]. With fundamentally different underlying architectures, MEMS-based storage promises seek times ten times faster than hard disks, storage densities ten times greater, and power consumption one to two orders of magnitude lower. It can provide several to tens of gigabyte non-volatile storage in a single chip as small as a quarter, with low entry cost, low shock sensitivity, and potentially high embedded computing power. It is also expected to be more reliable than hard disks thanks to its architectures, miniature structures, and manufacture processes [5, 11, 21]. For all of these reasons, MEMS-based storage is an appealing next-generation storage technology, especially in mobile computing applications where power, size, and reliability are important.

Due to the limited capacity of a single MEMS device, storage systems require such devices, also the corresponding connection components, one to two orders of magnitude more than disks to meet their capacity requirements. This can significantly undermine system reliability and increase system costs. Thus, we propose to integrate multiple MEMS devices, organized as RAID-5, into a *MEMS storage enclosure*. MEMS storage enclosures are the building block of MEMS-based storage systems, whose role is exactly the same as disks'—providing reliable persistent storage.

Besides data and parity devices, a MEMS enclosure also has the flexibility to contain several on-line spares thanks to the low entry cost, small physical size, and low power consumption of MEMS devices. Because of the high bandwidth and limited capacity of MEMS devices, data on a failed device can be recovered to on-line spares in minutes, which significantly reduces the risk of data loss during data reconstruction and improves system reliability. The enclosure notifies the host system, the maintenance personnel and/or the end users when it runs out of spares. It can be either upgraded by a new enclosure or replenished with new spares to increase its life time, depending upon users' decisions.

We find that the Mean Time To Failure *(MTTF)* of MEMS enclosures is smaller than that for disks unless we aggressively replace spares, but the probability of data loss during the economic lifetime (3–5 years) of an enclosure is lower than that of a disk. Thus we conclude that MTTF is not an appropriate metric. Failed MEMS devices can also be replaced when necessary or desired. A simple preventive

[†]Supported by the National Science Foundation under grant number CCR-0073509 and the Institute for Scientific Computation Research at Lawrence Livermore National Laboratory under grant number SC-20010378.

[‡]Supported in part by an SCU-internal IBM research grant.

maintenance strategy can make MEMS enclosures highly reliable.

For space reasons, we do not consider in this paper the internal organization of data within a single MEMS device. Just like disks, MEMS devices will deal internally with such failures as media defect and random bit error. In addition, MEMS will also confront new failure modes such as tip failures. We did some preliminary research on using advanced Error Control Coding (ECC) to deal with these problems. As a result, a MEMS device will present a similar abstraction as modern disk drives do in that a stored block of data will be retrieved with very high probability and that data corruption is exceedingly unlikely.

## 2. MEMS-based Storage

A MEMS-based storage device is comprised of two main components: groups of probe tips called *tip arrays* that are used to access data on a movable, non-rotating *media sled*. In a modern disk drive, data is accessed by means of an arm that seeks in one dimension above a rotating platter. In a MEMS device, the entire media sled is positioned in the *x* and *y* directions by electrostatic forces while the heads remain stationary. Another major difference between a MEMS storage device and a disk is that a MEMS device can activate multiple tips at the same time. Data can then be striped across multiple tips, allowing a considerable amount of parallelism. However, the power and heat considerations limit the number of probe tips that can be active simultaneously; it is estimated that 200 to 2000 probes will actually be active at once.

Figure 1 illustrates the low level data layout of a MEMS storage device. The media sled is logically broken into non-overlapping *tip regions*, defined by the area that is accessible by a single tip, approximately 2500 by 2500 bits in size. It is limited by the maximum dimension of the sled movement. Each tip in the MEMS device can only read data in its own tip region. The smallest unit of data in a MEMS storage device is called a *tip sector*. Each tip sector, identified by the tuple $\langle x,y,tip \rangle$, has its own servo information for positioning. The set of bits accessible to simultaneously active tips with the same *x* coordinate is called a *tip track*, and the set of all bits (under all tips) with the same *x* coordinate is referred to as a *cylinder*. Also, a set of concurrently accessible tip sectors is grouped as a *logical sector*. For faster access, logical blocks can be striped across logical sectors.

Table 1 summarizes the physical parameters of the MEMS-based storage device used in our research, based on the predicted characteristics of the second generation of MEMS-based storage [21]. While the exact reliability numbers depend upon the details of that specification, the techniques themselves do not.

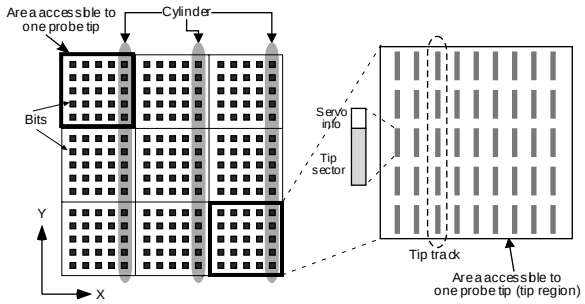



**Figure 1. Data layout on a MEMS device.**

**Table 1. Default MEMS-based storage device parameters.**

| | |
|---|---|
| Per-sled capacity | 3.2 GB |
| Maximum throughput | 76 MB/s |
| Number of tips | 6400 |
| Maximum concurrent tips | 1280 |

## 3. Related Work

Although MEMS-based storage is still in its infancy and no public literature is available on its reliability, the MEMS technology itself has played important roles in automotive industries, medical technologies, communications, and so on [1]. Among them, Digital Micromirror Devices (DMD) is a commercial MEMS-based digital imaging technology developed by Texas Instruments (TI). Douglass [7, 8] reported and estimated its Mean Time Between Failure *(MTBF)* was 650,000 hours.

System designers have long tried to build reliable storage systems. RAID (Redundant Arrays of Independent Disks) [6, 10, 19] have been used for many years to improve both system reliability and performance. Traditionally, system designers were more concerned with system performance than reliability during data reconstruction. Menon and Mattson and Thomasian [15, 16, 26] evaluated the performance of dedicated sparing [9], distributed sparing [16], and parity sparing [20] under the normal and data recovery modes of RAID systems. Muntz and Lui [17] proposed that a disk array of *n* disks be declustered by grouping the blocks in the disk array into reliability sets of size *g* and analyzed its performance under failure recovery.

Disk manufacturers are widely using S.M.A.R.T (Self-Monitoring Analysis and Reporting Technology) to recognize conditions that indicate a drive failure and provide sufficient warning before an actual failure occurs [2, 23, 14]. The preventive replacement strategy used in our MEMS storage enclosures can be viewed as a coarse-grained device failure predictor.

## 4. Reliable Storage Building Blocks—MEMS Storage Enclosures

Thanks to its non-volatility, MEMS-based storage can replace or complement disks in storage systems. In general, disks have much higher storage capacities than MEMS storage devices, whose expected capacity is 3–10 gigabyte [21]. For instance, the capacities of state-of-the-art hard disks range from 18–300 GB for server SCSI disks, 40–400 GB for desktop IDE disks, and 20–80 GB for laptop IDE disks [13, 24], reported in August 2004. Thus, storage systems require MEMS devices more than hard disks by one to two orders of magnitude to meet their capacity requirements. Correspondingly, more connection components, *i.e.* buses and interfaces, are also needed. These can significantly undermine system reliability and increase system costs.

The advance of the magnetic disk technology is not well-balanced. The increase in disk capacity noticeably outpaces the increase in bandwidth [12]. Thus disk rebuild times are becoming longer, during which a subsequent disk failure (or a series of subsequent disk failures) can result in data loss. Because MEMS devices are expected to have at least as high, if not higher, bandwidths as hard disks and their capacities are limited, device rebuild times are significantly shorter for MEMS devices than for disks, which can in turn reduce the vulnerability window length thus improve system reliability. The small physical size, low power consumption, and relatively low entry cost of MEMS devices make it flexible to add on-line spare MEMS devices in storage systems to improve their reliability.

Because of the reliability and cost concerns, we believe that multiple MEMS devices should be integrated into one MEMS storage enclosure under one interface and organized as RAID-5. We choose RAID-5 as the data redundancy scheme because of its reliability, space efficiency, and wide acceptance.

The role of MEMS storage enclosures in storage systems is exactly the same as disks'—providing reliable persistent storage. A controller manages MEMS devices in an enclosure and exposes a linear storage space through the interface. MEMS enclosures are the basic building block of MEMS-based storage systems, just as disks in disk-based systems.

Besides data and parity devices, a MEMS enclosure also has several on-line spare MEMS devices to improve its overall reliability, durability, and economy. The controller is able to detect device failures in seconds or minutes. As long as there are spare devices, data recovery can start immediately without replacement ordering and human interferences, which significantly reduces the window of data vulnerability and the chances of human errors thus improves the MEMS enclosure reliability.

When an enclosure runs out of spares, it can notify the host system, the maintenance personnel and/or the end users through signals. For example, a red / amber / green LED combination might inform a laptop user of the state of the MEMS. The enclosure can be either upgraded by a new enclosure or replenished with new spares to increase its life time, depending upon users' decisions. Although an enclosure without spares can still tolerate one more failure thanks to the RAID-5 organization, a preventive replacement/repair strategy can be still desirable because it can significantly improve the system reliability.

Adding on-line spares can reduce maintenance costs because maintenance for such enclosures can be less frequent. It can also improve the enclosure durability because an enclosure can tolerate several device failures in its economic lifetime.

## 5. Reliability of MEMS Storage Enclosures

MEMS storage enclosures are internally organized as RAID-5 with on-line spares. Researchers traditionally approximate the lifetimes of RAID-5 systems as exponential and use Mean Time To Failure *(MTTF)* to describe their reliability [6, 10, 19]. This approximation is accurate enough because the lifetimes of the system components are also modeled as exponential and failed components can be replaced in time, *i.e.* the system is repairable. Thus, with failed device replacement, MEMS enclosures share similar reliability characteristics with RAID-5 systems and their lifetimes can also be modeled as exponential.

However, without failed device replacement, the lifetimes of MEMS enclosures can be viewed as two stages: the reliable stage with spares and the unreliable stage without spares. When it still has spare devices, a MEMS enclosure can be as reliable as RAID-5 systems with very short rebuild times; when spares run out, the enclosure becomes unreliable because any two successive device failures can result in data loss. Thus the lifetimes of MEMS enclosures without replacement cannot be simply modeled as exponential.

### 5.1. Reliability without Replacement

We first study the reliability of MEMS storage enclosures with dedicated spares in an idealistic but simple situation. The spare devices do not participate in request services during normal operations. We only consider the reliability of MEMS devices; other components in the enclosures are perfect. Failed MEMS devices are not replaced.

We assume that a MEMS enclosure contains 19 data, one parity, and $k$ dedicated spare devices. The user-visible capacity of the enclosure is 60 GB because the capacity of a single MEMS device is 3.2 GB (see Table 1). We assume

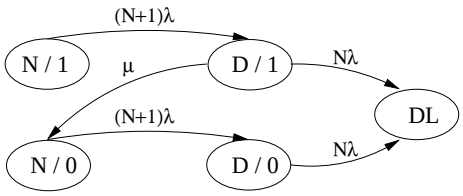



**Figure 2. Markov model for a MEMS storage enclosure with $N$ data and one parity devices and one dedicated spare. The enclosure can be in three modes: normal (N), degraded (D), and data loss (DL). We assume that MEMS lifetimes are independent and exponential with mean $MTTF_{mems} = 1/\lambda$. Recovery times of failed devices are also independent and exponential with $MTTR_{mems} = 1/\mu$. The numbers (0 or 1) in the figure indicate how many spares the enclosure still has.**

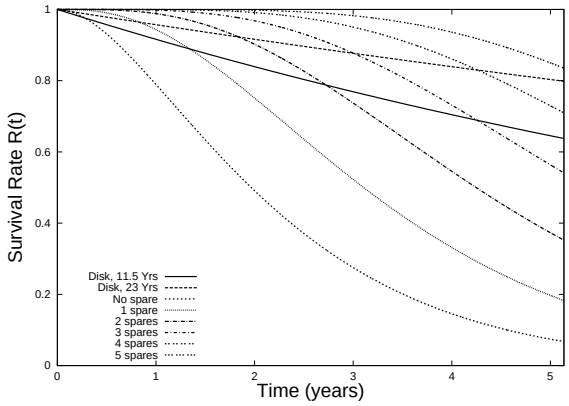



**Figure 3. Survival rates of MEMS storage enclosures using dedicated sparing without failed device replacement in five years.**

data rebuild times from a failed device to an on-line spare are exponential with $MTTR_{mems} = 0.25$ hour (Mean Time To Repair). It is a very conservative estimation, considering the high bandwidth (76 MB/s) and relatively small capacity (3.2 GB) of MEMS: we only use less than 5% of the device bandwidth for data recovery.

Unfortunately there is no data on the reliability of MEMS-based storage devices because they are still being developed and not commercially available yet, although researchers and engineers of MEMS-based storage expect that MEMS storage devices are more reliable than disks [4, 11]. Only limited literatures on the reliability of microelectromechanical systems are publicly available today. For instance, Digital Micromirror Devices (DMD), a commercialized MEMS-based digital imaging technology, have Mean Time Between Failure *(MTBF)* of 650,000 hours (74 years) [7, 8].

For simplicity, we assume that MEMS-based storage devices have exponential lifetimes with the mean of 200,000 hours (23 years). For the purpose of comparison, we assume the lifetimes of commodity disks and "better" disks are also exponential with means of 100,000 and 200,000 hours, respectively. While the exact reliability numbers depend upon these assumptions, the techniques themselves do not.

Figure 2 gives the Markov model for a MEMS storage enclosure with $N$ data and one parity devices and one dedicated spare. By using a simple method described in [10], *MTTF* of such systems with $s$ dedicated spares can be approximated as

$$MTTF \quad \doteq \quad \frac{s+1}{(N+1)\lambda} + \frac{1}{N\lambda},$$

where $1/\lambda$ is the average lifetimes of MEMS devices. Thus, *MTTF* of MEMS enclosures with zero to five spares are 2.3, 3.5, 4.6, 5.8, 6.9, and 8.1 years respectively, which are surprisingly low.

Although with low *MTTF*, MEMS enclosures with several spares can be more reliable than single devices with *MTTF* as high as 200,000 hours (23 years), even though the enclosures are not repaired in their economic lifetimes. Figure 3 illustrates the survival rates of MEMS enclosures without repairs. The survival rate $R(t)$ of an individual system is defined as the probability that the system survives for any lifetime $t$ given that it is initially operational [25]:

$$R(t) \quad = \quad \Pr[lifetime > t \mid initially\ operational].$$

Figure 3 indicates that with 3–5 dedicated on-line spares a MEMS enclosure is more reliable than a single device with *MTTF* of 23 years in the first 3–5 years, even without repairing the enclosure. For instance, the probability of data loss due to the failure of a MEMS enclosure with five spares in the first three years is 1.75%, much better than 12.31% of a single disk with *MTTF* of 23 years. However, when it runs out of spares, the enclosure becomes unreliable and the probabilities of data loss due to enclosure failure in one month and one year are 0.235% and 21.06%, respectively.

Typically, a disk with an exponential lifetime has a flat S-shaped survival rate curve. In comparison, MEMS enclosures, which also have S-shaped survival rate curves, achieve higher survival rates in the beginning but then rather suddenly fall under the survival rate of a disk, as shown in Figure 3. Thus, even though a MEMS enclosure might have a smaller *MTTF*, its survival rate for 1, 2, ... years can be significantly better than that of a disk. Basically, the enclosure survival rate follows no longer an exponential distribution, but a Weibull-type distribution. Economic lifetimes (3-5 years) are much smaller than component *MTTF* ($> 10$ years), which explains the seemingly paradoxical situation

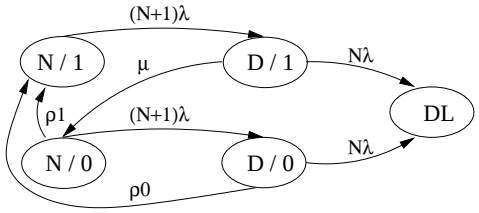



**Figure 4. Markov model for a MEMS storage enclosure with $N$ data and one parity devices and one dedicated spare. The times for replacing failed devices are independent and exponential. The mandatory and preventive replacement rates are $\rho_0$ and $\rho_1$, respectively.**

that enclosures with lower *MTTF* are more reliable than a disk drive.

Fortunately, the host system, the maintenance personnel, and/or the end users can notice when an enclosure enters its unreliable stage then schedule a repair in time. Note that MEMS enclosures are only the building block of storage systems. All the data on a "unhealthy" enclosure can be replicated to an on-line spare enclosure within one hour, assuming 17 MB/s bandwidth consumption, which is only 1.2% of the aggregated bandwidth of the MEMS enclosure.

### 5.2. Reliability with Replacement

When they run out of spares, MEMS enclosures can ask for maintenance services. There are two replacement strategies in MEMS enclosures: the preventive strategy schedules replacement right after spares run out and the mandatory strategy schedules replacement only when the enclosures operate in the degraded RAID-5 mode without any spares. Figure 4 shows the Markov model for a MEMS enclosure with $N$ data and one parity devices and one dedicated spare with replacement, where $\rho_0$ and $\rho_1$ are the mandatory and preventive replacement rates, respectively. We assume that the times to replace failed devices are independent and exponential.

Preventive replacement can significantly improve the reliability of MEMS enclosures because they can still tolerate one more failure during the replacement time, typically in days or weeks, thanks to their internal RAID-5 organization. Mandatory/nonpreventive replacement postpones enclosure repairs as late as possible so it can reduce the maintenance frequency during the lifetime of the enclosures. However, nonpreventive replacement makes users exposed to higher risks of data loss or unavailability.

Figure 5 shows *MTTF* of MEMS storage enclosures with different numbers of dedicated spares under different replacement strategies and replacement rates, ranging from one day to three months. We fix the number of data devices ($N$ = 19) and the average data recovery time to on-line spares ($1/\mu$ = 15 minutes). Clearly, on-line spares with preventive replacement can dramatically increase *MTTF* of MEMS enclosures, about one to two orders of magnitudes higher than *MTTF* of enclosures without on-line spares, under the same replacement rate. Without preventive replacement, the reliability improvement by on-line spares is less impressive.

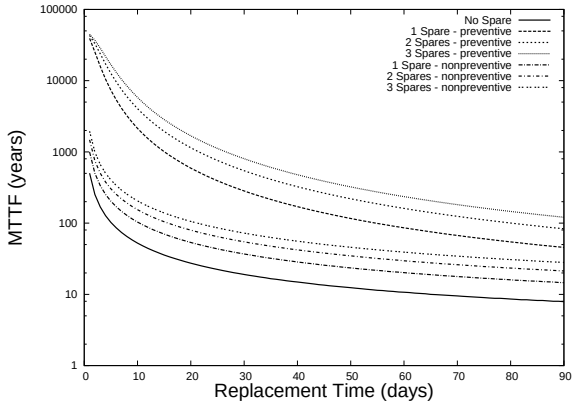


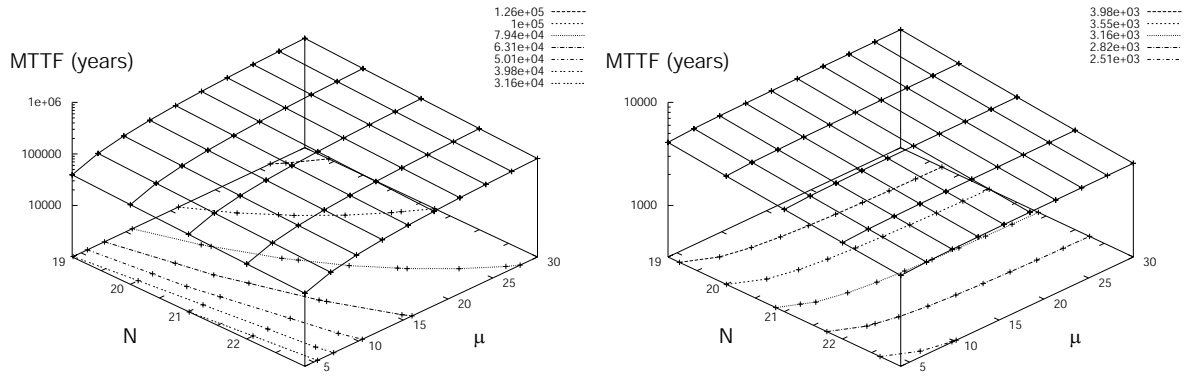

**Figure 5. *MTTF* of MEMS storage enclosures using dedicated sparing under different replacement strategies and replacement rates (as represented as $\rho_0$ and $\rho_1$ in Figure 4). We set $\rho_0 = \rho_1 > 0$ in the preventive replacement strategy and $\rho_0 > 0$ and $\rho_1 = 0$ in the mandatory/nonpreventive replacement strategy.**

The reliability *(MTTF)* of MEMS enclosures is heavily dependent on how fast failed devices can be replaced: when the average replacement time increases from one day to one month, *MTTF* of enclosures can drop by one to two orders of magnitudes. Compared to nonpreventive replacement, preventive replacement can reduce replacement urgency under the same reliability requirement, as shown in Figure 5.

The number of active data devices $N$ and the average data recovery rate to on-line spares $\mu$ also have impacts on MEMS enclosure reliability. Figures 6(a) and 6(b) show *MTTF* of MEMS enclosures, with one dedicated spare and using preventive replacement, as a function of $N$ and $\mu$ given that the average replacement times are one day and one week, respectively. We vary $N$ from 19 to 23 and $\mu$ from 4 (15 minutes) to 30 (2 minutes), which are reasonable ranges for MEMS enclosures under consideration.

In general, *MTTF* decreases with the increase of $N$ and the decrease of $\mu$. Note that the changes of *MTTF* under the specified ranges of $N$ and $\mu$ are within four to five times. Thus, $N$ and $\mu$ have less profound impacts on *MTTF* than the average device replacement rates, $\rho_0$ and $\rho_1$, as shown in Figure 5.

Given a relatively large replacement time (one week on




(a) One-day average replacement time

(b) One-week average replacement time

**Figure 6. Contour figures for *MTTF* of MEMS storage enclosures with $N$ data devices and one dedicated spare under failed device replacement times of (a) one day and (b) one week. Preventive replacement is assumed. Different data recovery rates to the on-line spare, $\mu$, are also examined.**

| MEMS 1 | MEMS 2 | MEMS 3 | MEMS 4 | MEMS 5 | MEMS 6 |
|---|---|---|---|---|---|
| D1 | D2 | D3 | D4 | P1 | S1 |
| D6 | D7 | D8 | P2 | S2 | D5 |
| D11 | D12 | P3 | S3 | D9 | D10 |
| D16 | P4 | S4 | D13 | D14 | D15 |
| P5 | S5 | D17 | D18 | D19 | D20 |
| S6 | D21 | D22 | D23 | D24 | P6 |

(a) Before any failure

| MEMS 1 | MEMS 2 | MEMS 3 | MEMS 5 | MEMS 6 |
|---|---|---|---|---|
| D1 | D2 | D3 | P1 | D4 |
| D6 | D7 | D8 | P2 | D5 |
| D11 | D12 | P3 | D9 | D10 |
| D16 | P4 | D13 | D14 | D15 |
| P5 | D18 | D17 | D19 | D20 |
| D23 | D21 | D22 | D24 | P6 |

(b) After MEMS 4 fails

**Figure 7. Distributed sparing.**

average), *MTTF* mostly relies on $N$ instead of $\mu$, as shown in Figure 6(b). When replacement tends to be postponed, the risk of data loss during the short data reconstruction time is neglectable, compared to the risk of data loss during the long replacement period. However, the data reconstruction rate becomes more relevant when the replacement time becomes shorter (one day on average), as shown in Figure 6(a).

### 5.3. Reliability of Distributed Sparing

Spare storage in MEMS enclosures can also be organized in a distributed fashion. In distributed sparing [16], client data, parity data, and spare space are evenly distributed on all of the devices in the enclosure. This technique can provide better performance than dedicated sparing in the normal and data reconstruction modes [15, 16, 26]. Compared to dedicated sparing, distributed sparing needs to reconstruct less data than dedicated sparing and its data reconstruction can be processed in parallel from and to all devices, avoiding the serialized reconstruction problem in dedicated sparing. Thus distributed sparing can reduce data reconstruction times thus reduce the window of vulnerability and the risk of data loss. However, distributed sparing utilizes more devices, which may undermine the overall enclosure reliability. Figure 7 gives a well-known layout of distributed sparing.

Figure 8 shows the Markov model for a MEMS enclosure with $N$ devices using distributed sparing. Because the MEMS enclosure generally stays in the data reconstruction modes for a very short time, we can safely merge the reconstruction modes to the normal modes by adding transitions directly from the normal modes to the data loss mode with small probabilities, $N\lambda q_j$, to simplifying our calculations. The probability of data loss during the data reconstruction time $t_r$ when $j-1$ devices still survive, $q_j$, is $1-e^{-(j-1)\lambda t_r} \doteq (j-1)\lambda t_r$, where $t_r$ is always no larger than its counterpart in dedicated sparing.

Distributed sparing and dedicated sparing can provide comparable or almost identical reliability to the MEMS enclosure configurations under examination. Figure 9 compares *MTTF* of MEMS storage enclosures using either dedicated or distributed sparing with different numbers of spares under different replacement rates, ranging from one day to three months. The user-visible storage is 60 GB, which is equivalent to the total storage of 19 MEMS devices. We set the data recovery rates to on-line spares in distributed sparing higher than those in dedicated sparing.

Distributed sparing requires less time to reconstruct data to on-line spares, which can improve reliability; on the other hand, it involves more active devices, which can undermine reliability. These two effects can balance each other, as shown in Figure 9 and Figures 6(a) and 6(b). In particular, dedicated sparing and distributed sparing provide almost

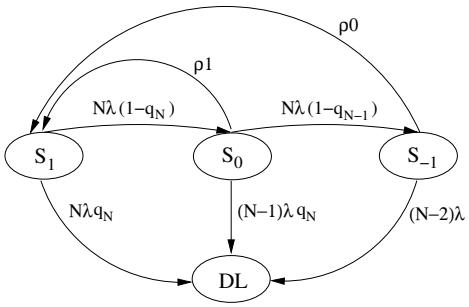



**Figure 8. Markov model for a MEMS storage enclosure with $N$ devices using distributed sparing. We assume that MEMS lifetimes are independent and exponential with mean $MTTF_{mems} = 1/\lambda$. We distinguish states $S_k, S_{k-1}, \cdots, S_0, S_{-1}$ where the index indicates the number of virtual spare devices left. State $S_{-1}$ is the state in which the parity data is already lost. $DL$ is the state of data loss. The probability of data loss during data reconstruction when $j-1$ devices still survive is $q_j$. Failed component replacements are independent and exponential. The mandatory and preventive replacement rates are $\rho_0$ and $\rho_1$, respectively.**

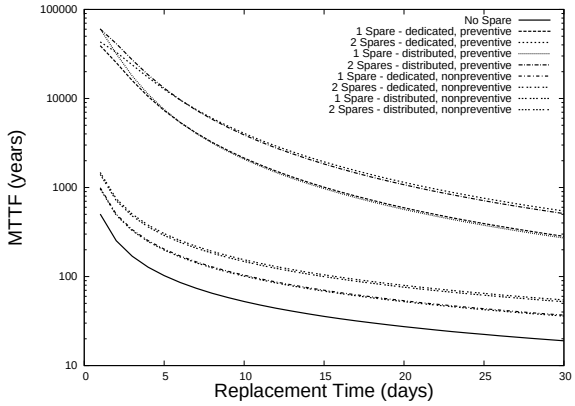



**Figure 9. *MTTF* of MEMS storage enclosures using either dedicated or distributed sparing under different replacement rates, as represented as $\rho_0$ and $\rho_1$ in Figures 4 and 8. We set $\rho_0 = \rho_1 > 0$ in the preventive replacement strategy and $\rho_0 > 0$ and $\rho_1 = 0$ in the mandatory/nonpreventive replacement strategy.**

identical MTTF to MEMS enclosures under the mandatory/nonpreventive replacement strategy. Typically, the average device replacement time is in days or weeks and the average data reconstruction time to on-line spares is in minutes. Thus, without preventive replacement the risk of data loss during device replacement is much higher than the risk during data reconstruction.

Although distributed sparing has shorter data reconstruction times, it has no significant impact on MEMS enclosure reliability. When preventive replacement is employed, the risk of data loss during data reconstruction is comparable to the risk under fast replacement because the replacement time is short and the enclosures can tolerate one more failure during the replacement period. Thus, distributed sparing provides better reliability than dedicated sparing only under this situation, as shown in Figure 9.

### 5.4. Other Issues on MEMS Storage Enclosure Reliability

In Sections 5.1–5.3, we assumed only MEMS devices in MEMS storage enclosures can fail and other components are perfect. Failures of MEMS devices are also assumed to be independent. In reality, data loss can be caused by a variety of reasons, such as correlated device failures, shared component failures, system crashes, unrecoverable bit errors, and so on. We simply follow the failure analysis in [6, 22] and present the results here.

Like disks, MEMS device failures tend to be correlated due to common environmental and manufacturing factors. Also alive devices in an enclosure after the initial failure generally have to service much heavier workloads than usual due to both external requests and internal data reconstruction requests. For simplicity, we follow the assumption that each successive device failure is ten times more likely than the previous failure until the failed device has been reconstructed [6]. Under this assumption, *MTTF* of MEMS enclosures with or without preventive replacement drops by 9–10 times. In particular, preventive replacement is more desirable than mandatory/nonpreventive replacement because it can still provide high reliability without urgent repairs, as shown in Figure 5. For instance, the probabilities of data loss due to a double device failure in the first three year for a MEMS enclosure with one dedicated spare, using either preventive or nonpreventive replacement, under a one-week-average repair rate are 0.65% and 14.54%, respectively.

We assume that all MEMS devices in an enclosure are attached to common power and data strings. Then the probabilities of data loss due to string and controller failures in the first three years are 0.52% and 2.59% respectively, assuming the power string has *MTTF* of $5 \times 10^6$ hours (571 years) and the RAID-5 controller has *MTTF* of $10^6$ hours (114 years) [22]. It suggests that the controller more likely results in data loss than MEMS devices although it is much more reliable than a single MEMS device.

Reliability estimations, following the approaches in [6],

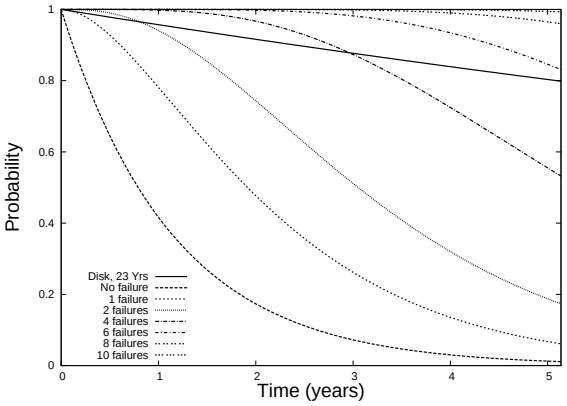



**Figure 10. Probabilities that a MEMS storage enclosure has up to $k$ failures during $(0,t]$.**

illustrate that the probabilities of data loss due to uncorrectable bit errors and system crashes followed by a MEMS device failure in the first three years are 0.14% and 0.18%.

In summary, as the basic building block of storage systems, MEMS enclosures can be much more reliable than disks, even with the considerations of possile data loss due to double failures, shared component failures, system crashes, and unrecoverable bit errors. Calculations show that the probability of data loss of a MEMS enclosure with one spare under preventive replacement in the first three year is 4.0%, which is significantly lower than 12.3% of a single disk with *MTTF* of 23 years. Note that MEMS enclosures are only the building block of storage systems and higher levels of redundancy should be provided to protect against data loss by MEMS-enclosure-based systems, as in disk-based systems.

## 6. Durability of MEMS Storage Enclosures

In MEMS storage enclosures, failed devices tend to be replaced as late as possible or even not replaced during the economic lifetimes of the enclosures to minimize maintenance costs and human interferences. This strategy raises questions on the durability of MEMS enclosures: how long can they work without repairs? How many times do they need repairing in the first 3–5 years? How do different replacement policies affect the maintenance frequency?

Again we consider a MEMS enclosure with 19 data, one parity, and $k$ dedicated spare devices. For simplicity, we assume that data reconstruction to on-line spares completes instantaneously as one device fails. Let $p_n(t)$ be the probability that exactly $n$ MEMS devices in the enclosure have failed during the period of $(0,t]$. As discussed in [3],

$$p_n(t) = e^{-\lambda_N t}(\lambda_N t)^n \frac{1}{n!}, \tag{1}$$

where $\lambda_N = N\lambda$, $N$ is the number of data and parity devices in the enclosure and $1/\lambda$ is *MTTF* of MEMS devices, which is assumed to be 23 years in our study. Thus, the probability that a MEMS enclosure confronts up to $k$ failures during the period of $(0,t]$ is

$$\begin{aligned} P_k(t) &= \sum_{n=0}^{n=k} p_n(t) \\ &= \sum_{n=0}^{n=k} e^{-\lambda_N t}(\lambda_N t)^n \frac{1}{n!}. \end{aligned} \tag{2}$$

In other words, the enclosure can survive after time $t$ with the probability of $P_k(t)$ as long as it can tolerate up to $k$ failures. Figure 10 illustrates the probabilities that up to $k$ failures occur in a MEMS enclosure during $(0,t]$.

Without any repairs, a MEMS enclosure with $k$ spares can tolerate up to $k+1$ failures in its lifetime. With $m$ repairs ($m \geq 1$), the enclosure can tolerate up to $k \times (m+1)$ failures under preventive replacement and $(k+1) \times (m+1)$ failures under mandatory/nonpreventive replacement before the $(m+1)th$ repair is scheduled. Here we assume enclosure repairs can be completed instantaneously because we are interested in how many times an enclosure has to be repaired during its economic lifetime, instead of its reliability.

For comparison, the probabilities that a disk with *MTTF* of 23 years can survive for more than one, three, and five years are 95.7%, 87.7%, and 80.3%, respectively. A MEMS enclosure with two spares has the chance of 98.8% to survive for one year without repair. The probability that an enclosure with five spares can survive for five years without repair is 84.6%. The chance that an enclosure with three spares under preventive replacement requires more than one repair during five years is 15.4%; instead, the chance for the same enclosure under nonpreventive replacement is only 3.5%. Adding one more spare can further reduce these probabilities to 3.5% and 0.6%, respectively. Obviously, preventive replacement trades more maintenance services for higher reliability, compared to mandatory replacement.

Figure 10 is almost identical to Figure 3 because the average data reconstruction time to on-line spares is very short in reality. Thus the assumption of immediate data recovery in Figure 10 is quite accurate in calculating the reliability of MEMS enclosures without repairs. Therefore, we can quickly get the approximation of the survival rates of MEMS enclosures without repairs by using Equation 2, without solving messy ordinary differential equations.

## 7. Concluding Remarks

Although MEMS-based storage is expected to be more reliable than disks, injudicious usage of such devices can result in significant reliability degradation in computer systems. We propose to pack multiple MEMS devices,

along with on-line spares, into one MEMS storage enclosure, which is the basic building block in storage systems. MEMS enclosures can be more reliable than disks even without repairs in their economic lifetimes, say 3–5 years. Furthermore, a simple preventive replacement policy can make MEMS enclosures highly reliable with *MTTF* of more than 1,000 years. We also find that dedicated sparing and distributed sparing have no appreciable difference on MEMS enclosure reliability.

## Acknowledgments

We are grateful to Ethan L. Miller, Qin Xin, and Feng Wang for their invaluable discussions. We also thank other SSRC members for their support.

## References

[1] AllAboutMEMS.com. MEMS applications. http://www.allaboutmems.com/memsapplications.html, 2004.

[2] ATA SMART feature set commands. Small Form Factors Committee SFF-8035. http://www.t13.org.

[3] U. N. Bhat and G. K. Miller. *Elements of Applied Stochastic Processes*. John Wiley & Sons, Inc., New Jersey, 3rd edition, 2002.

[4] L. Carley, J. Bain, G. Fedder, D. Greve, D. Guillou, M. Lu, T. Mukherjee, S. Santhanam, L. Abelmann, and S. Min. Single-chip computers with microelectromechanical systems-based magnetic memory. *Journal of Applied Physics*, 87(9):6680–6685, May 2000.

[5] L. R. Carley, G. R. Ganger, and D. F. Nagle. MEMS-based integrated-circuit mass-storage systems. *Communications of the ACM*, 43(11):72–80, Nov. 2000.

[6] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys*, 26(2):145–185, June 1994.

[7] M. R. Douglass. Lifetime estimates and unique failure mechanisms of the digital micromirror device (DMD). In *Proceedings of the 36th IEEE International Reliability Physics Symposium*, pages 9–16, May 1998.

[8] M. R. Douglass. DMD reliability: a MEMS success story. In *Proceedings of SPIE*, volume 4980, pages 1–11, San Jose, CA, Jan. 2003. SPIE.

[9] R. H. Dunphy, Jr., R. Walsh, and J. H. Bowers. Disk drive memory. United States Patent 4,914,656, Apr. 1990.

[10] G. A. Gibson. *Redundant Disk Arrays: Reliable, Parallel Secondary Storage*. PhD thesis, University of California at Berkeley, 1990.

[11] J. L. Griffin, S. W. Schlosser, G. R. Ganger, and D. F. Nagle. Operating system management of MEMS-based storage devices. In *Proceedings of the 4th Symposium on Operating Systems Design and Implementation (OSDI)*, pages 227–242, San Diego, CA, Oct. 2000. USENIX Association.

[12] J. L. Hennessy and D. A. Patterson. *Computer Architecture—A Quantitative Approach*. Morgan Kaufmann Publishers, 3rd edition, 2003.

[13] Hitachi Global Storage Technologies. Hitachi Disc Product Datasheets. http://www.hgst.com/, August 2004.

[14] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan. Improved disk-drive failure warnings. *IEEE Transactions on Reliability*, 51(3):350–357, 2002.

[15] J. Menon and D. Mattson. Comparison of sparing alternatives for disk arrays. In *Proceedings of the 19th International Symposium on Computer Architecture*, pages 318–329, Queensland, Australia, May 1992. ACM Press.

[16] J. Menon and D. Mattson. Distributed sparing in disk arrays. In *Proceedings of Compcon '92*, pages 410–416, Feb. 1992.

[17] R. R. Muntz and J. C. S. Lui. Performance analysis of disk arrays under failure. In *Proceedings of the 16th Conference on Very Large Databases (VLDB)*, pages 162–173, Brisbane, Queensland, Australia, 1990. Morgan Kaufmann.

[18] Nanochip Inc. Nanochip: Array nanoprobe mass storage IC. Nanochip web site, at http://www.nanochip.com/preshand.pdf, 1999.

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

[20] A. L. N. Reddy and P. Banerjee. Gracefully degradable disk arrays. In *Proceedings of the 21rd International Symposium on Fault-Tolerant Computing (FTCS '91)*, pages 401–408, Montreal, Canada, June 1991. IEEE Computer Society Press.

[21] S. W. Schlosser, J. L. Griffin, D. F. Nagle, and G. R. Ganger. Designing computer systems with MEMS-based storage. In *Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 1–12, Cambridge, MA, Nov. 2000. ACM Press.

[22] M. Schulze, G. Gibson, R. Katz, and D. Patterson. How reliable is a RAID? In *Proceedings of Compcon '89*, pages 118–123. IEEE, Mar. 1989.

[23] SCSI "Mode sense" code "Failure prediction threshold exceeded". American National Standards Institute. http://www.t10.org.

[24] Seagate Technology, Inc. Seagate Disc Product Datasheets. http://www.seagate.com/products/datasheet/, August 2004.

[25] D. P. Siewiorek and R. S. Swarz. *Reliable Computer Systems Design and Evaluation*. The Digital Press, 2nd edition, 1992.

[26] A. Thomasian and J. Menon. RAID5 performance with distributed sparing. *IEEE Transactions on Parallel and Distributed Systems*, 8(6):640–657, June 1997.

[27] J. W. Toigo. Avoiding a data crunch – A decade away: Atomic resolution storage. *Scientific American*, 282(5):58–74, May 2000.

[28] P. Vettiger, M. Despont, U. Drechsler, U. Urig, W. Aberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, and G. Binnig. The "Millipede"—More than one thousand tips for future AFM data storage. *IBM Journal of Research and Development*, 44(3):323–340, 2000.