Title

Modeling the Future of Archival Storage Systems

Author

Byron, James Lewis

Publication Date

2022

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MODELING THE FUTURE OF ARCHIVAL STORAGE SYSTEMS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

James Lewis Byron

September 2022

The Dissertation of James Lewis Byron
is approved:

Prof. Ethan L. Miller, Chair

Prof. Darrell D. E. Long

Prof. Erez Zadok

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Modeling the Future of Archival Storage Systems

by

James Lewis Byron

The importance of archival storage has increased with the growing demand for cost-effective long term data storage. Existing storage technologies like tape, hard disk drives, and solid state disks can meet today's demand for archival capacity and performance, but recently their pace of development has decelerated below their historical norms. In order to meet the ever-growing demand for archival storage, novel storage technologies are in development that will contend for their own place in archival storage systems.

The long-term viability of existing and novel storage technologies depends on how well suited they are for the demands of archival systems in the long term. We have created a simulation model to examine the relationships between different candidate storage technologies for archival systems and to define each technology's role within the competitive archival storage market over the long term. We enumerate the economic advantages of each storage technology relative to capacity, reliability, and workload for archival storage systems, and we describe how each technology can preserve its advantages and defend is position relative to other candidate archival storage technologies. We argue that novel storage technologies like synthetic DNA and glass will deliver decisive advantages in terms of cost, reliability, and scalability. Novel technologies will therefore dominate existing and traditional storage technologies for archival storage systems. Nevertheless, existing storage technologies will each retain a niche role for archival storage systems that suits their design.

Dedicated to my dear friend Chloe.

Acknowledgments

It is hard to understate the contributions from my advisor, Professor Ethan Miller toward my achievement in completing this work. Without his many insights, criticisms, advice, encouragement, and patience through it all, this work would not have been possible. I am indebted to him for his support in every aspect of this work. I am also more than fortunate to have worked with Professor Darrell Long. Since the very beginning of my work in the Center for Research in Storage Systems, his brilliant insights have inspired me to strive for the highest standards of scientific precision and clarity. I am also honored to have collaborated with Professor Erez Zadok. His vast experience and thoughtful guidance have helped me to refine the ideas and presentation throughout this work. I would also like to acknowledge the industrial members and sponsors of the Center for Research in Storage Systems (CRSS) for their support and feedback on the work that contributed to this dissertation. Finally, my fellow students in the CRSS deserve my endless gratitude for their camaraderie, fellowship, and technical support.

Chapter 1

Introduction

The information age has yielded many advancements for the benefit of humanity across numerous domains. The discovery of medicines to cure or mitigate ailments that would, until recently, have shortened many lives emerges from the availability of information—particularly digital information—to process and yield novel insights. Advancements in banking and finance, communications, governance, natural resource management, and environmental sustainability all derive their innovation and progress in part from the computing resources of the information age.

The information age has emerged as one of the brightest in all of human history for its ability to solve important problems, and as its name implies, information is the fundamental element that unlocks so many important innovations. But the longevity of our advancements from the age of digital information must depend upon how long these advancements can last. Whereas the invention of the printing press supported progress through centuries by helping to record and preserve information for the long term, the innovations that rely on digital information depend upon much less durable and long-lived storage media than ink and paper. Without proper care for the longevity and durability of the information from our digital civilization, the information age could, with the loss of data, fall from one of the brightest times in history to one of the darkest.

The preservation of digital information over the long term remains one of the great challenges within computer science. Data archiving is then one of the most crucial, albeit less glamorous, of projects for the digital age. Already today there exist numerous technologies that are candidates to help us solve the challenges of long-term data preservation. Such technologies have been in development for many decades, but as each of these technologies has only a limited lifespan, it remains essential to proactively guard, maintain, and upgrade the archival storage systems that store our digital treasures.

The long-term maintenance of archival storage systems depends upon the willingness of its stewards to continue their work, and the cost of maintaining an archive will undoubtedly affect how willing and able are the stewards to continue their work. Storage engineers and the stewards of archival systems must not only preserve data over the long term, but they must also preserve the software and other tools that were needed to read the data in the archival system. Moreover, as technology continues to progress, the growing demand for archival storage has motivated engineers to propose new storage technologies that can minimize the long-term cost of maintenance while meeting the ever-growing demand for archival storage capacity.

The demand for copious and economical archival storage has driven storage engineers to continuously improve existing storage technologies, helping them to meet the archival storage demand with ever-increasing capacity and performance. As new storage technologies are under development to meet and exceed the demands of tomorrow's archival systems, existing storage technologies must either adapt to compete with emerging technologies or, alternatively, relegate themselves to niches within the archival storage market. I hypothesize that device reliability will not limit the viability of storage technologies for long-term archival storage since other aspects of storage devices like capacity and performance are more significant factors in the cost of archival systems. Furthermore, I hypothesize that emerging storage technologies will outclass their legacy peers in terms of cost, capacity, reliability, and performance to dominate the archival storage market, and existing storage technologies will struggle to compete with each other for a diminishing role within the archival storage market.

In order to evaluate each candidate storage technology, we have designed a simulation model for archival storage systems that incorporates both existing and novel storage technologies. We explore a variety of possibilities, both in terms of features for each storage technology and design requirements for the simulated archival storage system, to quantify the differences between each technology and demonstrate its long-term viability relative to other storage technologies. We evaluate each storage technology in a monolithic archival storage system, using only one type of storage during each simulation, and we do so in order to measure the differences between storage technologies rather than their proportional representation within a hybrid storage system. Finally, we compile and present our results in terms of their support for long-term reliability and workloads in archival storage systems.

The following chapters include the background to our work, including a definition of terms, a description of each storage technology, and related work. Next we describe the operation of our simulation model and the parameters that we use to characterize each storage technology. We present results on the cost of reliability in archival storage systems using each candidate storage technology. Then we present the results of evaluating the cost of using each storage technology within archival systems to serve various archival workloads, and finally we conclude with a chapter that compares the predictions of our simulator with those from another model that compared two technologies using a simpler and static model.

Chapter 2

Background

The demand for long-term archival storage has given rise to numerous technologies and previous work relating to its development. Here we present relevant details about the technologies and state of related research on archival storage.

2.1 Definition of Archival Storage

Our discussion of technologies for future archival storage must begin with a definition of how we define archival storage. Simply put, we define an archive to be a data storage system that preserves data which may or may not be accessed over an extended period of time, typically measured in years or decades. Archival data may survive longer than the lifetime of the devices on which it is stored, and future generations of storage engineers will migrate archival data from older storage devices to new ones in order to preserve archival data for posterity. Our definition implies that archival storage systems must write all of their stored data at least once, but they may read only part of the data that they store over the lifetime of the storage system. Additionally, archival storage systems tend to last a long time, and as such, they benefit from reliable storage technologies that require minimal maintenance over time. Archival systems also tend to be capacious to store vast amounts of data in a cost-efficient manner. Thus, our working definition for archival storage highlights three important attributes: how often its users access the archived data, the length of time over which the archive stores data, and what data the archive stores.

We do not consider performance to be a central aspect of our definition of archival storage systems. Performance can be an important aspect of archival storage systems if, for instance, the users of the archival storage system must access the archived data with high throughput, however infrequently they may do so [92]. Primary storage systems designed around solid state disks or hard disk drives may also deliver high performance to their users [130]. Performance in terms of throughput, then, may relate to both primary and archival storage systems, and we do not consider throughput to be an essential aspect of our definition of archival storage.

Archival systems, like all storage systems, naturally require all data to be written to them at least once; however, archived data may seldom or never be read once written into the archival storage system [5,90]. The relationship between writes and reads in archival storage has been called *Write-Once-Read-Maybe*, which suggests that users read data in the archival system with some probability that is less than a certainty. The write-once description also suggests that archived data is written *exactly* once; however, previous work has shown that the write-once description should constitute a lower bound for write operations per data object rather than a strict proscription since users may sometimes update data within an archival storage system [5]. We summarize the relationship of write and read operations in an archival storage system as that of a storage system workload with write operations proportional to the data added to the archival storage system and read operations matching the behavior of the archive's users. The workload for a storage system, that is, the combined read and write operations over its lifespan, help to identify it as an archival storage system if some data in the storage system is read rarely or never. However, such a description of archival systems in terms of their workload does not preclude the possibility that primary or other non-archival storage systems may match our description; nevertheless, data that matches the workload characteristic *Write-Once-Read-Many* may indeed become good candidates for data eligible to transfer into an archival storage system. We also further enhance our definition for archival storage with description of the duration of time over which the storage system preserves its data.

Archival data storage systems tend to preserve their data over long periods of time. Prior work analyzing archival storage systems has often utilized archival workload traces or records that cover only a relatively short period of time [4, 62], an unfortunate side effect of the limited availability of archival storage traces from sources that have operated archival storage systems over long periods of time. Archival data may persist for decades, centuries, or longer as the technology and external factors allow [53, 75]. Storage technology, as it ages, becomes obsolete or disused, sometimes falling into disrepair, and sometimes being forgotten altogether. Disused, obsolete, and faulty storage devices present the predictable challenge of proving to be difficult or impossible to recover their data as the technologies used to encode and write the data have long since retired from active use [130]. Such challenges are incidental to the nature of long-term archival storage because they are side effects of the infrequent access patterns of archival data as discussed above, the unrealized economic value of archived data, and the cost and potential difficulty of preserving the data. Archival storage systems therefore must survive not only technical challenges like long-term reliability, capacity, and performance requirements, but also provide a stable and usable platform that future generations of storage engineers will find tractable and economically sustainable to preserve in perpetuity. The latter reemphasizes our motivation for this work as we seek to explore possible trends of future archival technology and support the preservation of archival data.

Beside characteristics of our definition for archival storage relating to its workload and duration, the large capacity of archival storage is an important and typical characteristic. As with previous features that distinguish archival storage from other storage systems, having a large capacity is not necessarily indicative of an archival storage system. Nevertheless, capacity—in particular, capacity for a minimal total system cost—is an important design feature for archival storage systems. We take this assumption because, as argued in previous work, the ongoing maintenance cost of an archival storage system significantly affects its long-term viability [20, 110]. Archival storage systems may store backups of primary storage, infrequently-accessed sensory data such as surveillance footage, scientific data, or other large structured or unstructured data [117]. Typically, archival storage is deployed for such datasets because of its large capacity and relatively low unit cost for storage, and the datasets that occupy archival storage systems typify the largest datasets that any storage tier preserves. Large capacity is therefore part of our definition for archival storage, since archival storage systems generally offer large capacity with lower cost than other storage systems.

With our working definition in place for archival storage and how its characteristics contribute to its definition, we proceed now to our discussion of different storage technologies that may construct archival storage systems.

2.2 Archival Storage Technologies

Archival storage systems utilize any of a variety of technologies that each have their own unique technical characteristics that affect their practical and economic viability within archival storage systems. The technologies that we consider include archival tape, optical disc, hard disk, solid state disk, archival glass, and synthetic DNA. Here we present each storage technology and highlight details that affect their utility within archival storage systems.

2.2.1 Traditional and Non-Traditional Technologies

Traditional archival storage technologies store data by mapping physical positions on a tape, disc, or platter to memory addresses. Accessing data on such storage technologies involves first spinning a tape cartridge, moving an optical laser, or actuating a hard disk head over a rotating platter to the physical position that corresponds to an address where the needed data resides. Each action necessarily involves a physical action that precedes any electronic retrieval of data. We describe tape, optical disc, and hard disk as traditional archival storage technologies because of their long history of availability and frequent adoption in long-term storage systems, either for personal or enterprise use. The constraints of physical operations needed to access different memory locations on traditional archival storage media necessarily limits the speed at which they can access, read, and write data. Furthermore, the effects of physical actuation in storage devices affects certain device characteristics more than others.

The need to physically actuate traditional archival storage media affects their performance and features asymmetrically. Hard disks and tape drives, for instance, have consistently improved their read and write throughput alongside their storage capacity over time [22, 37, 79] since throughput for such traditional storage technologies is a function of each device's areal bit density and the speed at which the storage medium moves. Latency, on the other hand, does not improve with areal bit density since latency varies by the physical size of the medium, the limits of the device’s tolerance for acceleration, and the speed at which the device rotates or moves [141]. Tape devices may deliver longer latency even as their throughput increases if manufacturers increase the length of the tape within each cartridge to increase capacity [64]. Hard disk drive latency remains largely unchanged because the speed of the platters and the speed at which the head moves over the platters do not improve as other aspects of hard disk technology develop [122]. With the exception of optical disc media [99], traditional storage technologies also require climate-controlled environments for storage in order to prevent damage to their moving parts and minimize degradation of their storage media [105, 129]. Non-traditional storage technologies for archival systems depart from some aspects of traditional technologies to overcome their associated limitations.

Traditional archival storage technologies—tape, hard disk, and optical disc—rely on mechanical operations to access data: move an arm, rotate a disk, or move a tape over a physical head in order to seek, record, or read data. The relationship between memory locations and data need not involve a physical operation. DRAM, for instance, stores data in a volatile solid-state medium with much faster performance than physically-actuated technologies. Solid state devices are those which have no moving parts; instead, they rely on transistors to electronically access different memory locations. The electronic basis of their operation gives the key to the better performance characteristics of solid-state devices in general, since electronic rather than physical operations dominate the device’s performance. Solid State Disks, like DRAM, have no moving parts within them, but SSDs offer the important benefit over DRAM of non-volatility or persistence after power failures. SSD storage has infrequently been applied to archival storage because of its higher cost relative to traditional archival technologies. We describe SSDs as non-traditional for archival systems due to their relatively recent emergence as a high-capacity storage option and their reputation for higher acquisition cost compared with traditional archival technologies. For the purposes of this work, we consider SSDs that use block-addressable interfaces such as those commonly found in consumer SSDs or in servers. The design and characteristics of Solid State Drives also highlights their benefits and constraints for long-term archival storage.

2.2.2 Archival Tape

Tape is perhaps the most dominant technology option for archival storage systems. Tape storage for digital information resembles those used for analog audio and video with cassette tapes, VHS tapes, and other similar technologies. Tape for audio and video purposes is today regarded as obsolete compared to their digital successors in optical disc, DVD, and digital downloads. Audio and video technologies for consumer-oriented consumption have become obsolete because, unlike other technologies that remain viable today, they have received no significant upgrades once they were released to the public. Thus, obsolete technologies identified themselves through their lack of regular upgrades and consequent technical stagnation. Archival tape for digital data, on the other hand, has received upgrades and improvements on a regular basis, including changes that require the replacement of tape cartridges and drives to remain up-to-date with new technology advancements [22].

The Linear Tape-Open (LTO) Consortium publishes the specification and roadmap for current and future generations of archival tape technology [76]. The LTO format is the product of a joint effort between International Business Machines (IBM), Hewlett-Packard Enterprise (HPE), and Quantum, among others. As such, the LTO tape standard and its contributing technologies draw on the efforts and resources of developers and engineers throughout the world, and with its distributed development and manufacturing model comes the benefits of having many manufacturers and interested parties for tape technology. Other storage technologies like hard disk drives and optical disc have suffered from a consolidation in their markets due to the rising capital costs of continued development, and although the same cost inflation applies to tape development, cooperation between vendors through the LTO Consortium may mitigate such risks.

Although the LTO Consortium has been organized to minimize the risks of consolidation that other competing technologies have endured, there remain risks to tape's development that even the cooperative framework has not solved. A legal dispute that began in 2016 between Sony and Fujifilm—the only two manufacturers for critical components in the latest generation of tape cartridges—delayed the availability of the needed capacity and performance upgrade of LTO-8 by nearly two years [83, 84]. The LTO-8 legal dispute demonstrates risks to archival storage systems that exist outside the scope of technical developments and economic considerations. Such risks are beyond the scope of our work, but they are nevertheless critical to consider when evaluating the potential for disruptions that may affect the viability of any given storage technology.

An archival storage system that utilizes tape for storage includes numerous components to read, write, store, and deliver data. The tape cartridge, which stores the archival data, can last for long periods of time without losing data, assuming that the surrounding environment remains temperature and humidity-controlled over time [3,51]. The need for constant environmental controls is one constraint for tape which some—bot not all—storage technologies also demand [7, 10, 59, 88]. Each archival tape cartridge contains a length of tape film that has a magnetic material bonded to its surface. The magnetic material can degrade over time if the storage environment becomes too hot or humid for long periods of time, but in most cases, manufacturers of archival tape promise a 30 year lifespan for cartridges [51]. The tape inside each cartridge can be hundreds or thousands of meters in length, and since the data on the tape is stored magnetically along the length of the tape, random access operations that require seeking for data can take a minute on average to perform [51, 64]. Decreasing the latency of tape would require either reducing the length of the tape or increasing the speed at which tape moves through a tape drive, but the physical constraints of tape drives and the demand for high capacity cartridges inhibits improvements to the latency of archival tape [49, 64, 132]. Nevertheless, what tape lacks in terms of latency it makes up to some extent in the throughput it offers during read and write operations. Tape drives today can read and write data at 300 MB per second, and each generation of tape drive promises to increase its throughput [21, 78].

Archival tape develops by way of discrete generations that feature incrementally more capacity and throughput than the last generation [78]. The plan for future generations—the *roadmap*—signals to storage engineers and planners alike that archival tape will offer predictable capacity and throughput increases with which they can plan for future demand growth. Each generation of tape drive also offers reverse compatibility with one or two older generations of tape cartridge media. Reverse compatibility helps to ensure that data stored on older generations of tape remain accessible for years after its respective generation of tape drive has become obsolete. The length of time between each generation of tape technology is typically two or three years, so reverse compatibility can offer four to six additional years of accessibility for older tape generations [22, 77].

The pattern of development for tape drives and media offers flexibility for designers of archival storage systems that use tape-based storage; however, each generation of tape cartridge will eventually become unreadable by new generations of tape drive. For this reason, storage system engineers must plan for the obsolescence and retirement of older tape technology by migrating the data stored on older generations to newer, higher capacity, and more performant generations of tape. Doing so will also help to prevent a scenario where older tape drives that fail mechanically over time will not result in a situation where data stored on old tape cartridges have no working tape drive that can read them, a condition that would result in the loss of data from the archive. Storage engineers therefore plan migrations between generations, and they benefit from the increasing capacity and performance of newer generations of tape as time passes. Migration between generations of tape, as with all storage technologies, requires ongoing management and capital investment [20]. In addition to their need for periodic migration, tape-based archival storage systems also require regular verification or scrubbing to identify any failing cartridges before they become unreadable [3, 95]. Scrubbing requires reading all the drives in the archive, a process that can sometimes require continuous operation by the drives in the archive.

Maintenance activities in tape-based archives like migration and scrubbing can sometimes demand constant operation by the tape drives in an archive. Tape-based archives, as with optical disc, scale capacity by adding more tape cartridges. While they add more tape media for capacity, tape-based archives do not necessarily require more tape drives that would have the effect of increasing the aggregate throughput of the archive as a whole. Tape-based archives therefore feature a many-to-one relationship between their storage media—tape cartridges—and their read and write mechanisms—tape drives. Incidentally, tape drives are much more expensive than tape media, so increasing the ratio between tape media and drives has the effect of reducing the per-byte cost of the archival system overall. The many-to-one structure of tape-based archival technology imposes performance constraints that technologies like hard disk drives and solid state disks do not suffer because, unlike tape media, each hard disk and solid state disk contains its own mechanism to read and write the data it stores [22, 49, 64]. Insofar as archival storage systems demand large capacity with minimal performance demands, tape can prove to be a viable technology; however, as the demand for archival throughput grows, the cost for archival tape increases rapidly with the required number of tape drives. Such design constraints are precisely why tape storage has proven to be so resilient as an archival storage technology. Considering our definition for archival systems based on capacity and workload from Section 2.1, the one-to-many design of tape archives has proven conducive for low-intensity workloads common to archival storage systems.

2.2.3 Solid State Disk

Solid State Disks (*SSDs*) have become increasingly popular in recent years as they have demonstrated their performance and reliability advantages relative to Hard Disk Drives [13]; however, their adoption in archival storage systems has been limited since their higher acquisition cost impairs their economic advantages over other storage media like Tape or HDD. Although SSD storage presents a higher acquisition cost than other storage media, some of its features may nevertheless prove to offer important advantages that could, under some use conditions, render it competitive with other storage technologies for archival systems.

The acquisition cost of SSD storage is a function of its cost of manufacturing and the areal bit density that each manufactured component delivers. Since NAND flash is the dominant technology within SSDs available today, we limit the scope of our discussion to the most prevalent existing SSD technology. There are three factors that contribute to the density of NAND flash SSDs: the feature size of individual cells of flash [16, 119], dimensional scaling by adding more layers to each wafer of flash [43, 48, 106, 119], and the number of bits per flash cell [43, 89]. Advancements such as charge-trap flash have helped to facilitate the continuation of each of the three modes for SSD scaling to further decrease the cost per byte of storage on SSDs [47]. Reducing the size of each cell is one possible mechanism for increasing the density of flash that involves neither adding more bits per cell nor adding additional layers of NAND flash.

NAND flash, like other types of solid-state electronics, relies on a lithographic process for manufacturing. Lithographic manufacturing requires appropriate machinery and supporting infrastructure that cost many millions or billions of dollars [2, 40, 135]. The lithographic manufacturing process also requires a template or mask that allows the machinery to project an image for the pattern of transistors, conductors, and other structures onto a silicon wafer. The lithographic mask can also cost many millions of dollars, depending on how many features and layers that the component requires [9, 135]. Each new generation of manufacturing technology and masks costs more than the last due to the greater complexity of the lithographic process technology, and the increasing capital costs must be passed on to customers through the manufactured products, thereby increasing the cost of the most advanced semiconductors. Over time, however, the cost for lithographic masks decreases for each lithographic node as the node ages [9], and other capital costs like machinery and factories become increasingly diluted with time through cost amortization. Solid-state electronics like NAND flash that are manufactured on older fabrication processes therefore become more economical with time while those utilizing the most advanced fabrication nodes become more expensive. NAND flash also suffers from additional constraints on the extent to which its fabrication processes scale to smaller lithographic nodes.

Reductions in the feature size of each flash cell has yielded enormous improvements to the economics of flash storage [16, 119]; however, just as the challenges grow for reducing the feature size of transistors in logic components like CPUs, so do the challenges and unwanted side effects for continued decreases in the feature size of flash [40]. Each flash cell consists of a capacitive layer and a semi-permeable gate layer of silicon that allows the capacitive layer to retain a charge [40]. The physical size of the capacitive and gate layers must shrink with each new down-scaling of the lithographic feature size for flash [40, 135]. With each feature now only a few nanometers in size, the capacitive layer can hold only a handful of electrons to represent different binary values. The decreasing number of electrons in the capacitive layer of each flash cell causes different bit-values to become closer together and to increase the risk of them overlapping [16, 40, 89]. The problem of overlapping and confused bit values in flash, which results in erroneous data, grows with the shrinking size of the lithography used. The semi-permeable gate layer, similarly, becomes less capable of isolating the charge within each cell as its size decreases on each lithographic down-scaling. The combined challenges of the capacitive and insulating layers becoming less effective as their sizes decrease effectively imposes a lower bound of size for flash lithography, and further improvements beyond such a boundary result in unreliable and short-lived flash memory [40]. Reliability and longevity, which are two aspects of storage that are vital to long-term archival storage, remain as essential attributes for SSD storage. Increasing the capacity of NAND flash must therefore arise from some other scaling mechanism.

Over the last decade, NAND flash has enjoyed large improvements to its areal bit density and economic viability through the introduction of three-dimensional scaling [40, 41]. 3D NAND flash expands on the two-dimensional size of traditional planer NAND flash by constructing multiple layers of flash, one on top of another. Three-dimensional NAND flash began with a handful of layers and has grown to 200 at the present time, and future generations of flash may introduce even more layers to allow capacity to grow [82, 93]. Adding more layers to flash is not analogous to adding more platters to hard disk drives as we discuss in Chapter 2.2.4 because, unlike hard disk drive platters, it requires no additional cost to add more layers to each flash wafer [82, 93]. Furthermore, unlike the miniaturization of features as described above, three-dimensional flash offers the advantage of utilizing older generations of lithographic technology, which helps to reduce capital costs [119]. Furthermore, 3D flash requires little additional fabrication time within the most expensive components of a semiconductor manufacturing facility compared with planer flash [119], and thus the improvements to capacity wrought by 3D flash come at little additional manufacturing costs. The ability of flash manufacturers will, to a large extent, control the costs of SSD-based storage in the long run, particularly given the practical limits of further lithographic scaling as described above. Beyond the addition of more layers within a three-dimensional flash memory, multi-bit storage within each cell may yet amplify the gains made in other aspects of NAND flash memory at the expense of durability.

The scaling of flash storage capacity through lithographic shrinking and three-dimensional designs have both increased capacity and decreased the cost-per-byte of storage on flash, but these require significant capital investment or improvements to fabrication techniques to maintain the pace of development for SSD capacity and cost. Multi-bit NAND flash, on the other hand, can utilize existing and relatively old fabrication technologies to scale the cost and capacity for flash [43, 89]. Multi-bit flash scales capacity at the expense of device performance and durability, depending on the number of bits that share each flash cell [24, 25, 36, 41, 80,

Table 2.1: Endurance of Multi-Bit Flash

Name	Bits/cell	Charges	Capacity	Endurance [36, 80]
Single-level (SLC)	1	2	–	100000
Multi-level (MLC)	2	4	2.0×	10000
Triple-level (TLC)	3	8	1.5×	3000
Quad-level (QLC)	4	16	1.33×	1000
Penta-level (PLC)	5	32	1.25×	NA
Hexa-level (HLC)	6	64	1.2×	NA

127]. Different storage applications therefore favor different types of multi-bit flash.

In order to help distinguish between different flash technologies and their appropriate uses, we rely on the common nomenclature that identifies how many bits share each cell of flash. We present the names, bits-per-cell, and required number of charge states in Table 2.1.

Increases in the number of bits-per-cell of flash improves the areal density and decreases the cost-per-byte of storage on flash; however, numerous adverse effects diminish the advantages of flash over other technologies as the number of bits in each cell increases. Increasing the bits per cell causes the flash to become slower at writing and reading data. As noted in Table 2.1, each increase to the number of bits per cell results in a doubling of charge states that the NAND flash controller must distinguish during program and erase operations, and the greater number of charge states effects a narrowing of each voltage level that corresponds to the bit values in a given flash cell [89]. The exponential growth of change states requires the NAND controller and its supporting firmware to become ever-more sensitive to minute differences in cell voltage, since a small voltage change that would be insignificant for SLC or MLC flash can indeed change the value stored within a TLC or QLC flash cell. QLC flash and its descendants also suffer from potential read disturbance effects, where reading data in one block can incidentally change values stored in an adjacent block and result in unexpected data loss. For these reasons, adding more bits per cell of flash imposes both a small additional cost for more advanced controller circuitry and more sophisticated firmware that can manage the complex effects of storing more bits in each cell [89]. Increasing the number of bits per cell also reduces the overall endurance of the flash memory since, given the exponential growth of charge levels needed to store each additional bit, any degradation in the ability of the flash to retain a certain voltage in each cell increases the likelihood of data loss; adding more bits per cell narrows the differences between the voltages corresponding to each bit value [31, 81, 121]. The undesirable side effects of increasing the number of bits per cell impose a practical limit on the ability of SSD memory to reliably store data over the long term in an archival setting. SSDs, though less sensitive to temperature changes than other technologies like tape or hard disk, presents its own constraints to environmental factors that can impact its reliability over time.

SSDs, though less sensitive to temperature than other technologies like HDD or Tape, still remains sensitive to operating and storage temperatures; however, the characteristics of NAND flash offer certain advantages and disadvantages compared with other technologies with respect to temperature. Temperature can have unique effects on the operation of SSDs. Elevated operating temperature positively affects write performance and, in the case of multi-bit NAND, reduces read disturbs between adjacent blocks [32]. Elevated temperature reduces the effectiveness of the insulating layer within each cell of flash, and the lower barrier to changing the charge of each flash cell enables write operations to complete more quickly. Read operations complete with a lower probability of changing the charge states of other adjacent cells. However, while elevated temperature has certain positive effects by temporarily changing the properties of the flash memory, these same properties have undesirable effects in the context of archival storage.

Technologies for archival storage systems must be evaluated with respect to their characteristics that relate to the definition of archival storage as described in Section 2.1. Any benefits for short-term performance due to high operating temperature prove unimportant for archival storage if those benefits come at the expense of reliable long-term data preservation. As the workload of archival storage systems fits a write-once-read-maybe profile, the marginal performance improvement of the already-fast SSD technology from high operating temperatures improves the candidacy of SSDs for archival storage only slightly, yet since the primary effect of high temperatures on SSDs is to increase the possibility of electron leakage from each flash cell, we contend that high temperatures exhibit a deleterious overall effect on the long-term preservation of data in SSD-based archival systems.

While high temperatures have some positive effects on the performance of SSDs, the viability of SSDs for long-term data preservation declines as temperature increases due to the physical effects of temperatures on data retention in NAND flash memory [32, 106]. The relationship between temperature and data retention informs how SSDs perform in archival systems and what environmental constraints must apply to archival systems that employ NAND flash-based storage technology [32, 106]. Environmental constraints like temperature can directly affect the ability of storage technologies like flash to preserve data over long periods of time, and the relationship between temperature and data retention depends on the reaction rate of the storage medium within SSDs [32, 69, 106]. The Arrhenius equation defines a generalized model of the temperature dependence of reaction rates for a given material relative to a known baseline temperature for that material [69, 96]. The Arrhenius equation is used to calculate the acceleration factor (*AF*) for temperature-dependent reactions such as those within SSDs that cause charge leakage and data loss over time [69, 96]. We use the Arrhenius equation to extrapolate the reaction rate of NAND flash relative to its known baseline reaction rate at a certain temperature. The reaction rate of NAND flash indicates how quickly the charge state of each flash cell leaks toward zero volts, which constitutes data loss in SSD technology [69, 96]. The Arrhenius equation is defined as:

$$\text{AF} = e^{\frac{-E_a}{k} \times (1/T_2 - 1/T_1)}, \tag{2.1}$$

where AF is the acceleration factor for the reaction rate, E_a is the activation energy intrinsic to the specific material, k is Boltzmann's constant, and T_1 and T_2 are the baseline temperature and long-term storage temperature in degrees Kelvin, respectively [96]. The activation energy for SSDs, which expresses the amount of energy needed to leak change out of a flash cell, is given as 1.1Ev, and Boltzmann's constant is $8.623 \times 10^{-5}\text{eV}/°K$ [96]. The acceleration factor allows us to compare the expected longevity of data on an inactive SSD in storage under different temperature conditions, relative to the device's baseline data retention and temperature values as given by the manufacturer. Most SSDs promise to preserve data for 1 year at a storage temperature of $40°C$ or $313°K$ [96]. We use the acceleration factor, as calculated from the Arrhenius equation for a certain temperature, to scale up or down the length of time that we can expect the storage devices to preserve data by dividing the baseline duration of data on the storage medium by the calculated acceleration factor [96]. For some baseline retention period L, measured in years, we can use Equation 2.1 to calculate the expected data retention using the equation:

$$R_T = \frac{L}{\text{AF}_T}, \tag{2.2}$$

where R_T is the expected time for data preservation at a certain storage temperature [96]. To further illustrate the effect of storage temperature on SSD data

Table 2.2: Expected Retention Time vs. Storage Temperature for SSDs

$^{\circ}C$	$^{\circ}K$	AF	Exp. Retention Time
70	343	35.327	10 days
60	333	11.563	32 days
50	323	3.532	103 days
40 [96]	**313**	**1.000**	**365 days**
30	303	0.261	3.8 years
20	293	0.062	16.2 years
10	283	0.013	75.2 years
0	273	0.003	392.1 years

retention, we use Formula 2.1 to calculate and present various storage temperature values and their associated acceleration factors and expected retention times for data on SSDs in Table 2.2. The expected retention times show that higher temperatures can have a dramatic effect on retention times for data on SSDs. On the other hand, lower storage temperatures have an effect on SSDs that would be advantageous for long-term data preservation. We show the reference temperature of $40^{\circ}C$ in bold, and the reference temperature has an acceleration factor of 1 with an expected data retention time of 1 year [96].

Table 2.2 shows that relatively small increases in temperature can have a large and undesirable effect on data preservation for SSDs. Conversely, even modest decreases in ambient air temperature can significantly improve the data retention time for SSD technology, and a long data retention time can help to ensure that archival storage systems that utilize SSD technology avoid unwanted data loss [96].

2.2.4 Hard Disk

Hard Disk Drives (*HDDs*) trace their origins back decades [58] with the ever-present demand for capacious and reliable data storage at an economical total cost. Hard disk drives improve upon certain performance limitations of tape-based storage while preserving some of the cost and scalability advantages of tape. Hard disk thus delivers a compromise between tape with its cost and capacity advantages and DRAM or SSD with their speed advantage, albeit with numerous technological caveats that constrain the use and development of hard disk technology. HDDs consist of a group of platters attached to a high-speed electric spindle motor. One or more mechanically-operated arms, each with one head on the top and bottom for each platter surface, move over the spinning platters to read or write data [39]. The magnetic areas used for each bit of data are comprised of grains of magnetically reactive material that can be non-destructively read or written by the head as it moves over. The size and number of the magnetic grains, along with how closely they are packed together, determine the areal bit density of each platter. The platters are prepared during manufacturing to store data within logical blocks. As with tape media, hard disk drive platters are coated with a material that allows the head to read and write data onto the platters while minimizing the risk that data, once written to the platters, will change accidentally [112]. The magnetic coercivity of the material determines the strength of the magnetic field required to change one bit on the platter surface [70, 94], a factor that affects both the stability of data on the drive and limits the continued growth of areal bit density.

Hard disk capacity derives from the areal bit density of the hard disk platters and the number of platters within each drive. Adding more platters to each drive offers a straightforward and certain method for increasing drive capacity; however, adding additional platters to drives proves untenable in the long term. First, as each platter takes up some space within the hard drive, the physical size of the drive limit the number of platters that can be placed inside. The physical size of each drive must not change as its physical dimensions must exactly match the space made available to it within other hardware components. The fixed amount of space available for platters means that the number of platters cannot grow with enough consistency to meet the exponential demand for storage capacity. Furthermore, each additional platter incurs its own manufacturing costs and need for two additional heads. Each additional component and platter also presents a non-zero probability of failure, which, however small, increases the possibility that a small component defect will cause the entire drive to fail. Increasing the capacity of hard drives by adding more platters is thus an inferior approach for adding capacity compared with increasing the areal bit density of each platter.

The areal bit density of hard disk drive platters has grown consistently over time, allowing manufacturers to provide vastly improved capacity without increasing HDD prices [112]. Furthermore, and unlike the alternative of adding more platters, increasing the areal density of each platter delivers the added benefit of increasing the throughput of the drive. Areal bit density, measured as bits per mm^2, has grown consistently over time to deliver ever-greater throughput and capacity for hard disks at little to no additional unit cost per drive. In general, throughput of hard disk drives and optical discs increases with the square root of areal density since device throughput derives from the number of bits that can pass under the drive head each second [108, 114]. Tape media store data linearly rather than radialy, and therefore throughput on tape follows its own rules within each particular generation of tape technology [78]. Areal density includes the size of each bit on the hard disk platters and also how closely together are the tracks on the platters [108, 114]. Improvements in sensors and signal processing have long facilitated growth of areal density through the miniaturization of each bit on the hard drive platter. Looking into the future, further improvements to hard disk capacity and throughput encounter physical limitations that may limit further developments.

The areal bit density of hard disk drives has traditionally grown by reducing the size of each magnetic bit on the hard disk drive platter [114]. Today, the bits are of such a size that they increasingly encounter the physical constraints known collectively as the magnetic recording trilemma [70, 94, 98, 112]. As its name suggests, the magnetic recording trilemma consists of three components that are at odds with each other [70, 94, 98, 112]. To improve any one of them necessarily compromises at least one of the other two factors. The trilemma consists of readability, writability, and stability [70, 94, 98, 112]. Readability benefits from strong magnetic fields within the magnetic grains, but such strong magnetic fields are increasingly difficult to produce as the number of grains per bit decreases [94, 98]. Writability, on the other hand, becomes more challenging as the magnetic field needed within the grains, and by extension, the magnetic field needed to produce it, must grow stronger [94, 98]. Furthermore, the ever-decreasing size of each bit on the platter also proves more challenging for write operations because such small features require weaker magnetic fields and more finely targeted write operations that may not overcome the magnetic coercivity of the material [94, 98]. Magnetic coercivity therefore positively correlates with the ease and speed of read operations but negatively correlates with the ease and speed of write operations [94, 98]. Efforts to increase areal density by decreasing the size of each magnetic grain compromise the stability of the drive also by reducing its magnetic coercivity and its tolerance to temperature fluctuations [70, 112]. A hard disk drive platter with a high magnetic coercivity threshold may prove magnetically and thermally stable, but the drive's head will be unable to write data to the drive because its small size and the close proximity of magnetic grains will both limit the strength of the magnetic field that it can produce without creating excess heat or disturbing nearby bits [70, 112]. Significant research and development efforts have explored solutions to the trilemma with the promise of unlocking further areal bit density increases over the coming years [114].

Hard drive manufacturers have worked to increase the areal bit density of hard disk drives by escaping the magnetic trilemma. The magnetic stability of the grains on hard disk platters changes drastically with temperature by Formula 2.1. Technologies such as Heat-Assisted Magnetic Recording (HAMR) achieve both magnetic stability and writability by selecting an alternative magnetically-reactive material with a higher threshold of magnetic coercivity and by using a laser or directed energy to heat the platter before write operations. The action of heating the platter before write operations has the effect of reducing the threshold of magnetic coercivity for the magnetic grains, and then the write head with its small magnetic field can easily change the magnetic value of the bit. Read operations on heat-assisted drives occur without heating the platters, and since the magnetic coercivity of the grains is beyond the ability of the drive's heads to write data without first heating the platter, HAMR drives also offer thermal and magnetic stability under typical operating conditions [70, 94, 98, 112]. As areal density has increased consistently over time, hard disk drive features and limitations have also remained largely unchanged since the technology first emerged as an option for digital storage.

Hard disk drives characteristically require stable operating environments, including air conditioning in order to function reliably [7]. Hard disk drives also operate with a nominal power consumption between 5 and 10 watts [17], but as much as twice to three times this number when spinning up the platters from a standstill [1]. Storage systems that utilize hard disk drives can reduce power consumption by "spinning down" or powering off the spindle motor, thereby reducing the heat produced by the drives and potentially extending their lifetimes by reducing wear-and-tear [12, 125]. Drive startup is, however, a more demanding operation than normal operation due to the physical strain of moving the inert spindle motor and platters, and thus the action of powering down the hard drive to conserve energy and wear-and-tear must be balanced against the high power consumption, wear-and-tear, and delay of starting up the drive from a powered-off state. In addition to their power consumption and durability constraints, hard disk drives also present performance characteristics that distinguish them from other storage technologies.

The performance of hard disk drives lies between that of solid-state media like NAND flash and other magnetic storage like tape. Tape drives can require more than a minute to access data [50, 51], and SSDs using NAND flash can access data within microseconds [111]. Hard disk drives require several milliseconds to seek and access data on the platters [52, 115], assuming of course that the drives are already spinning. Although capacity and throughput of hard disk drives has improved dramatically over time, the latency of hard disk drives and the unit power consumption have improved only marginally. Reliability has also not improved significantly over time as manufacturers prefer to focus their efforts on capacity and throughput instead of long-term reliability [11]. The interest and investment of storage device manufacturers, which develop storage technologies to sell into the storage market, affect the availability and pace of development for storage technologies.

2.2.5 Optical Disc

Optical disc (*ODD*) technology consists essentially of a spinning plastic disc impregnated with a reflective material. The plastic disc spins on a spindle motor, and a mechanically operated arm moves across the surface of the spinning disc much in the way similar to that of hard disk drives. Also similar to hard disk drives, the throughput of optical disc varies with how quickly the disc spins and grows with the square root of its capacity. Optical discs utilize a laser to read and write information into the disc, and unlike HDD or SSD, optical disc media support only read operations once they have been written [118], a feature that finds frequent use within archival storage systems when data must not change once written into the archive. The information on the disc, which is stored optically in a chemically-stable nonvolatile material, can survive at least 50 years in a wide range of environmental conditions without experiencing degradation or data loss. Hence, optical disc technology is considered more resilient to environmental conditions than hard disk, solid state disk, or even tape technology [123, 124]. Optical disc also has a long history that has proven its reliability in a variety of applications [123, 124].

Optical disc drives (*ODD*) for data storage began with the introduction of compact discs (*CDs*) for digital audio in the early 1980s [107, 123, 124]. The CD medium resembles the much larger laser disc format for video media that was introduced a few years earlier for the home consumer market. Compact disc for audio and laser disc for video each served an emerging need for cost-effective, easily reproducible, capacious, and durable write-once storage media for the demands of digital media [107, 118, 123, 124]. At the same time, the increasing power and ubiquity of personal computers introduced a market for CDs to distribute software. Compact discs gave way to digital versatile disc (*DVD*) during the 1990s and eventually the Blu-Ray disc in the 2000s, each with greater capacity than the technology generation that came before [107, 123, 124]. Optical disc has proven unique among storage technologies insofar as older generations of the storage technology such as CDs and DVDs remain available for purchase and widely used notwithstanding the availability of the more modern and capable generations such as Blu-Ray disc. ODD technology also benefits from better reverse-compatibility than other technologies since even 40-year old CDs continue to operate in the latest generations of optical disc drives. Reverse-compatibility, the ability for new drives like optical disc drives and tape drives to read or write older optical disc or tape media, relates to technologies that feature a many-to-one relationship between recordable media and the drives used to read or write them. Optical disc and tape, each with separable media, each benefit from reverse-compatibility to some extent [77,107,123,124], but tape drives limit their reverse-compatibility to two or three generations of tape media as the technology continues to evolve. In the case of optical disc, the long lifetime of optical media—50 years—may deliver important benefits for archival storage applications as successive generations of drives continue to access even old media from decades earlier. Optical disc storage may then require less maintenance than other competing storage technologies as its reliability and reverse-compatibility minimize the need for data migration from older to new generations of disc technology. Although its reverse-compatibility may prove to be one of its principal advantages, optical disc may present challenges that impair its relevance within archival storage systems.

Other competing storage technologies have grown their capacity and throughput even faster than optical disc. Some have therefore questioned the relevance of optical disc as its industrial base narrows and its mind space within the storage industry shrinks [107]. As optical disc has become less common in personal computers in recent years, the pace of its development has, to a large extent, fallen behind that of other technologies and in particular behind the ever-growing demand for capacity from long-term storage systems [22]. Figure 2.1 illustrates the capacity of optical disc technology available over time compared with an example of capacity demand that grows with a 30% compound annual growth rate (*CAGR*).

Figure 2.1 shows the growing difference between the capacity of optical disc and that of an example demand for archival storage. The slower pace of ODD ca-

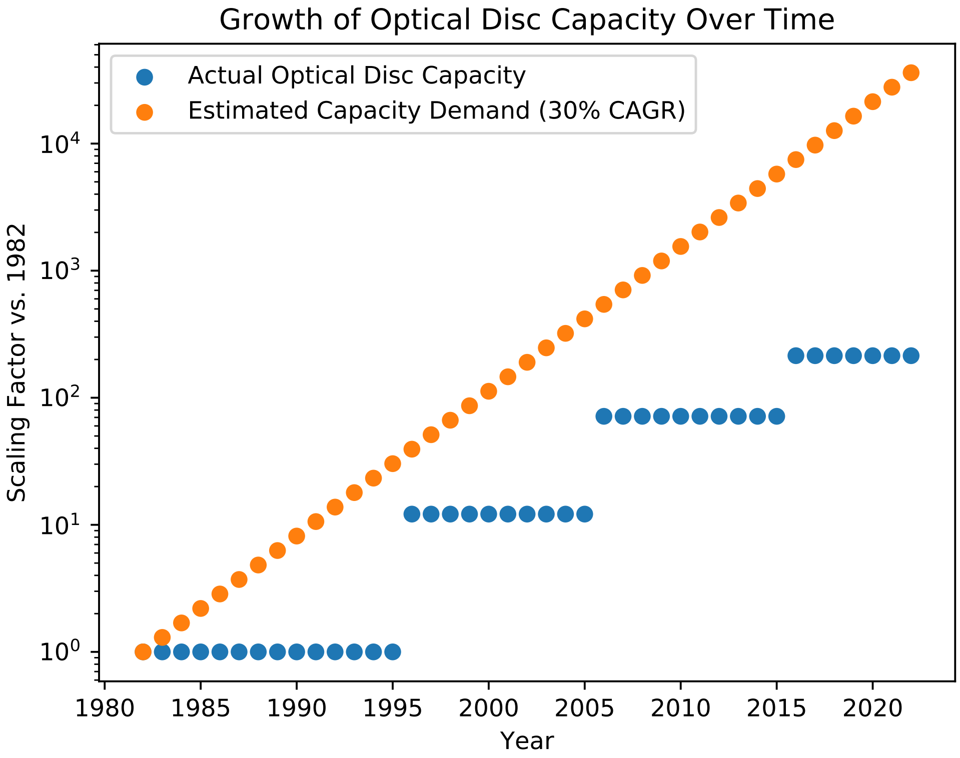

Figure 2.1: Each new generation of optical disc delivers more capacity than the one that came before. Here we show the capacity of each generation of optical disc normalized by the capacity of compact discs, the first generation of optical disc technology. We also show an normalized example of capacity demand for archival storage that grows 30% each year. The capacity of optical disc has grown only slowly over time, increasingly falling behind the scale of current capacity demand [14, 15, 19, 100, 124].

pacity growth, combined with decreasing interest in the technology from manufacturers, original equipment manufacturers, and consumers may lead optical disc to a marginal role in the storage market. Nevertheless, even as optical disc struggles to keep up with the inexorable growth of demand for archival storage capacity, the future of optical storage may yet provide an economical refuge for virtually limitless data. Recent attempts to increase the capacity of optical disc [85] could, if successful, help to reinstate optical disc technology as a leading contender to meet the demands of long-term archival storage. Optical disc also shares in common with archival glass the optical nature of its data storage.

2.2.6 Archival Glass

Archival glass improves upon the features of optical disc while significantly relaxing its constraints for capacity and cost. Archival glass bears numerous features in common with holographic data storage [54] as both technologies can function by repositioning a laser rather than moving the physical medium. Archival glass and holographic data storage also both utilize glass as a storage medium, but glass is a write-once medium by design, making it more suitable for archival storage systems [54]. Archival glass, like optical disc, stores data optically within a transparent medium, but unlike ODD, archival glass storage uses a stationary plate of glass and a movable laser to read and write data. Glass-based storage also reduces the cost of the medium compared with optical disc because data is stored within the glass crystal itself rather than in a chemically-stable medium impregnated within a plastic disc. Although its technical advantages over optical disc may prove significant, archival glass must also compete with other technologies for its place within the archival storage market, and since it remains today a prospective technology for archival storage systems, its viability for archival storage remains an open question.

Archival glass is a proprietary storage technology under development at Microsoft [8]. The glass storage medium resembles optical disc insofar as it minimizes the risk of bit rot and the attendant need for media scrubbing. Unlike optical disc, the active development of glass as a storage technology will produce growing interest from storage engineers, investment from companies like Microsoft, and eventually the technologies and products that are necessary to deliver archival glass as an economical storage medium. Microsoft expects its glass-based storage technology to become available for cloud deployments within the next decade [31].

The development of archival glass as a storage technology includes several components, each of which relate to other technologies that have, to some extent, previously proven themselves within other scientific applications. The storage medium—glass—exists abundantly in other forms within the earth's crust. For this reason, we expect the storage medium itself to remain the least expensive component of the archival glass storage medium. For the purposes of the storage system, the glass is stored in small rectangular plates. Each plate of glass may hold several dozen to hundreds of terabytes of data within an area of approximately 100 square centimeters. As with other storage technologies with removable media like optical disc and tape, a robot can mechanically transport the glass media between a collection or library of glass plates and the read or write drives in order to read or write data. The final component of the storage system must be the read and write drives that record and read data within the glass media. Archival glass may utilize an inexpensive laser paired with machine learning and computer vision algorithms to read data from the glass plates [8], but a more powerful laser must be used to write data into the glass material. For this reason, archival glass, unlike existing archival storage technologies, utilizes separate drives to read and write data onto the storage media [8]. The drives for reading and writing may also have different capital costs, power consumption, and throughput [8]. The exact costs of materials and components are presently unknown because, as noted earlier, archival glass remains today a prospective archival storage technology. Nevertheless, we project the costs of archival glass technology based on that of similar technologies.

2.2.7 Synthetic DNA

DNA has stored and communicated data since before the invention of the computer as even biological systems rely on DNA to represent the vast quantity of information that constitutes their genomes. Synthetic DNA has been investigated for long-term data storage for similar reasons [60], although today it remains a prospective archival storage technology for the somewhat distant future [31]. DNA is different from all other storage technologies—even archival glass—in that it could cost-effectively preserve orders of magnitude of information for far longer periods of time—a promise that other storage technologies strive endlessly to achieve. Instead of reacting to the storage demand of today and yesterday, synthetic DNA as a storage technology could provide enough capacity to suit the data needs of the distant future. Nevertheless, synthetic DNA as it exists today remains in its nascency.

Synthetic DNA, although it promises vast amounts of storage capacity, provides an undesirable performance profile for use as a storage technology. Any storage system must write all of its data at least once, but it may or may not read all of its data over time. Today, synthetic DNA technology provides much faster performance reading data than it does writing data [30,60]. For the purposes of DNA storage, read operations are called *sequencing*, and write operations are *synthesis*. DNA sequencing operations can take hours or days to complete [30], depending on how much data must be read. Synthesis operations can require one second per byte of data, which could require years or decades to complete a synthesis operation of a size commensurate with the capacity of DNA [30, 60]. Thus, in order to meet the demands of any conceivable archival storage system with its write-once-read-maybe workload profile, DNA storage must improve sequencing speeds by orders of magnitude and synthesis speed by many orders of magnitude. Still, development continues on DNA-based digital storage that may eventually render it a viable storage technology for archival storage.

Microsoft has demonstrated a end-to-end archival storage system which, although used to write only a small amount of data, demonstrated that DNA synthesis, preservation, and sequencers can be combined into a unified and coherent storage system [97, 126]. The demonstrated end-to-end storage system had a total cost of roughly $10,000. The costs of nucleic acids, the materials that must be synthesized to create DNA molecules, may also change over time as demand for storage induces supplies at an economical scale. Today, the costs of synthesizing DNA, equal to approximately $1 per million bits of data written, elevates the cost of synthetic DNA storage [143].

2.3 Related Work

Storage systems have been analyzed in other research in order to gather insights on their performance [5, 26, 27] and in order to predict their cost over time [22, 23, 45, 57].

2.3.1 Workload Characterization

A workload has been defined as “the set of all inputs that the system receives from its environment during any given period of time” [87]. A workload model therefore characterizes a real workload without repeating the the exact operations on the original data objects, and a workload model must stand in for real workloads for the purposes of system selection, performance tuning, or capacity planning [87]. Workload models can resemble real workloads in two ways: functional characterization and resource characterization [87]. Functional characterization models those aspects of the workload relating to programs, commands, and requests that constitute the workload while resource characterization models a workload's resource utilization such as CPU time, memory consumption, and total system accesses [87].

Several studies have analyzed workloads in cloud service provider and high-performance computing (HPC) systems [26, 27, 91, 103, 120]. HPC workloads can be generated from either real applications that utilize the HPC system or from synthetic workload generators [120]. Synthetic workload generators recreate a workload that mimics a real workload using either an empirical or analytical approach [87, 120]. Empirical synthetic workload generators replay the traces of a real workload to repeat operations that were previously executed on the target system [87, 120]. Alternatively, synthetic workload generators can analyze real workloads to create a mathematical model that can be replayed on a target system [87, 120]. The goal of creating an analytical model that accurately models a real workload can prove challenging yet essential for certain applications such as modeling and simulation [120].

Most recent research for workload analysis has focused on primary or high-performance storage systems [27,46,63,138,140,142] since, at least to some extent, researchers may prefer to work on projects relating to high-performance primary storage. The analysis of workloads in primary storage systems has informed their design and implementation, and the same importance of workload modeling applies to the design and implementation of archival storage systems [87].

Recognizing the need for empirical analysis of archival workloads, several studies have introduced techniques or results from analyzing workloads in archival systems. Wildani, et al. argued that the task of designing archival systems would benefit from more frequent and varied analysis of archival workloads akin to that which has been published for primary storage systems [136]. Furthermore, they showed that certain assumptions about archival workloads have remained true over time while others have not. In particular, assumptions about the frequency of data accesses vary from archive to archive, and this, in turn, has implications for archival system design.

Early work on archival workload analysis showed that recently archived data may be frequently accessed [61], thereby blurring the distinction between archival and primary storage. Later work compared traces from multiple archival systems, finding that the frequency of accesses far archival storage depends upon both internal factors such as user needs and maintenance and external factors such as search engine indexing [5]. The frequency of data updates also varies between different archival workloads. Archival workload analysis has also highlighted the possibility that storage technologies traditionally reserved for archival storage systems can, in some cases, better serve the needs for long-term archival storage [45, 68].

2.3.2 Archival Modeling

Several previous studies have offered modeling and simulation of archival storage systems to evaluate their characteristics over time. A comparison of hard disks and solid state disks for long-term archival storage concluded that SSDs may offer certain long-term advantages for archival storage due to their low power consumption, good reliability, and favorable capacity growth rate [45]. Other studies show that tape offers the lowest cost per byte of long-term storage [22, 23], but the limited performance of tape compared with other storage technologies leaves open the question of how different archival workloads may affect the total cost of ownership for an archive. Tape offers good reliability compared with hard disks; however, previous works have not considered prospective storage technologies like glass and synthetic DNA and how they may offer better features in terms of their capacity, reliability, or performance should they become available in the future.

Chapter 3

Methodology

We simulate archival storage systems that use either existing or prospective storage technologies, and we evaluate each storage technology by comparing it with others within the simulator. We compare existing storage technologies including tape, optical disc, hard disk drives, and solid state disks, and we extend our simulations to include the prospective storage technologies of archival glass and synthetic DNA that are currently in development. We base our models for existing storage technologies on the published data sheets from device manufacturers, and we infer parameters for prospective storage technologies from the publications that describe their design and performance. We use our simulation model and parameters for storage technologies to answer questions about each technology's place in the future of archival storage. The simulator uses parameters to perform each simulation for a period of 25 years, and each simulation returns a spreadsheet of information about the devices in the simulation, their performance, and the total cost of ownership for the simulated archival system.

3.1 Simulator Overview

The simulator takes as input parameters to describe storage technologies, environmental parameters, workload parameters, and general archive parameters. The parameters of the storage technologies include values such as device performance, capacity, power consumption, endurance, lifespan, and rates of development. Environmental parameters include the cost of electricity and its rate of annual increase. Workload parameters include the data storage of the archival system, the number of annual read and write operations, and the total throughput needed for the archival system. The archival parameters include values for the number of years to simulate, the resolution of the simulator, and the number of times to repeat each simulation. Each simulation uses the parameters together to conduct experiments and measure the cost, performance, or reliability of each storage technology.

3.1.1 Simulation Model

We evaluate archival storage systems on a range of metrics in order to conduct experiments and draw conclusions about the economic value of each storage technology. We utilize input parameters for the archival system, its workload, operating environment, and storage devices to calculate the total cost of ownership (*TCO*) over the time of the simulation. We design our simulation model to utilize several functions that each calculate the cost of one aspect maintaining an archival system over a known quantity of time. The functions for cost that we utilize include the cost of upgrading, maintaining, and powering the archival system. We summarize each cost function in Table 3.1.

The function for determining the cost of upgrading the archival system evaluates the current state of the simulated archival system in comparison with the requirements of the system. Upgrading the archival system becomes necessary as the demand upon the archival system in terms of storage capacity grows over time such that d increases for time t compared with time $t-1$, where d is the data demand for the archival system at time t. Thus, the function calculating the cost of upgrading the archival system returns a cost value that is proportional to the rate at which the archive's data grows and the amount of time elapsed between t and $t-1$.

We also consider the cost of the workload within the simulation model as part of the upgrade function. The workload of the simulated archival system controls the minimum performance needed by the storage system. Some storage technologies like tape feature separable media, effectively allowing archival systems with such technologies to scale capacity without necessarily increasing their performance capabilities. The workload of the simulation model ensures that the archival storage system can serve the demands of the workload as described in the workload parameters. The cost of the workload in the simulation model grows as the demands of the archive's workload increase for data d at time t relative to time $t-1$. We discuss the relationship between cost and archival workload in Chapter 5. Beside the cost of upgrading the the archival system, we also measure the cost of maintaining the archival system throughout the simulation time.

Maintenance costs on the archival system relate to the cost of maintaining the archival storage system to store data d with the reliability required of the archival system as defined in the archive's parameters. The failure of any individual device triggers maintenance of the archival system to recover data and replace the failed device with a new one. Storage devices that deliver high levels of reliability require less maintenance than devices with lower levels of reliability because fewer device failures result in fewer maintenance events. The cost of maintenance in the simulator is thus proportional to the quantity and reliability of the storage devices within the archival system. The reliability of storage devices, in turn, is proportional to their baseline reliability and also to each device's age since, as we discuss in Chapter 4, storage devices often become less reliable as they age. The reliability of individual storage devices can change as they age, particularly for device types like hard disk drives that have a limited lifetime after which their reliability decreases dramatically.

We also consider power cost as part of our simulation model. The cost of electricity in our simulation model depends upon both the amount of electricity consumed by the storage devices in the archival system and the cost of electricity at any point during the simulation. We discuss the cost of electricity in Chapter 3.3.1. We measure power consumption in kilowatt hours, and we evaluate power cost in terms of dollars per kilowatt hour. Power consumption for each device in our model depends on how long each device spends in each of three possible power states: active, idle, and standby. Active mode is when each device is actively reading or writing data into the storage medium. The idle mode is when the storage device is powered on, but it is not actively reading or writing data. Devices in the idle state may still have mechanical components like spindle motors powered on. Standby mode is when the device rests in its low power mode such that any mechanical components are turned off and only the device's minimal electronic components remain powered on. Each cost function—upgrade, maintenance, and power—takes as an input a time t which we evaluate discretely over a fixed slice of time called an epoch in the simulation. We calculate the total power cost of the archival system using the formula:

$$P_{d,t} = C_t \times \left(A_{d,t} + S_{d,t} + I_{d,t}\right), \tag{3.1}$$

where $P_{\mathrm{d,t}}$ is the calculated total power cost for data d at time t, $A_{\mathrm{d,t}}$ is the amount

Table 3.1: Summary of Functions to Calculate Cost

Cost Function	**Summary**
Upgrade	Calculate the current demand for the archival system in terms of data capacity demand and throughput. Calculate the current performance of storage devices available for the current epoch in the simulation. Purchase new devices as needed to increase the capacity or performance of the archival system while also maintaining the necessary parity for the simulation as defined in the configuration. Calculate the total cost of adding all devices to the archive. New devices will need to write data as they enter the archive, and their time spent writing data will increase the active power consumption of the system.
Maintain	Search for devices within the archival system that have experienced a failure during the current simulation epoch. Replace the failed devices with new devices, and compute the total cost of replacing them. New devices will also need to write data into them once they are installed into the archive, and the write operation will consume additional electricity.
Power	Compute the power utilization of all components within the archival system, including power consumption during active, idle, and standby modes. Calculate the current cost of electricity, and multiply the cost of electricity per kilowatt hour by the total kilowatt hours used by all components during the current simulation epoch.

of power consumed during active read and write operations in the archive over time t, $S_{\mathrm{d,t}}$ is the amount of power used during standby for devices in the archival system, and $I_{\mathrm{d,t}}$ is the amount of electricity used by idle devices within the archival system. C_t is the cost of electricity per kilowatt hour for the time t within the simulation.

We include as one of our goals the accuracy of our simulation model relative to real-world operations; however, each of the functions for calculating cost as part of our calculation for TCO requires some nonzero quantity of time over which to evaluate the archival system. Maintenance events, the purpose of which is to replace failed devices within the simulated archive, must evaluate the reliability of the archival system over some period of time, and each device within the archival system has some probability that it will fail over a certain amount of time. Therefore, in order to balance the goal of precision with the requirements of our cost functions, we allow the users of the simulator to configure the desired resolution of the simulator over which each separate action and cost function should evaluate time t. Our simulator supports values ranging from one second to one year of resolution, and for each of the experiments in the following chapters, we use one hour of resolution for the epoch of time t.

The ability of our simulation model to evaluate time with various degrees of granularity emerges from our goal to measure the effects of time and change on archival systems continuously. For this reason, we integrate for the total cost of ownership (*TCO*) with the continuous sum of all three cost functions over the duration of each simulation. We calculate the TCO of an archival storage system using the formula:

$$\mathrm{TCO}_{d,t} = \int_{t=1}^{T} \left(\sum_{f=1}^{F} \left(f(d,t) \right) \right) dt, \tag{3.2}$$

where $\mathrm{TCO}_{\mathrm{d,t}}$ is the total cost of ownership for the simulation up to time t and for the amount of data d, T is the total time of the simulation from beginning to end, and F is the set of functions f that return the cost of the archival system as described in Table 3.1. The upgrade and maintain cost functions calculate the number of devices needed to meet both the capacity requirement of d and also the number of devices required to serve the archival workload over the quantity of data d at time t.

3.1.2 The Simulation Model in Action

In this section we describe the operation of the simulator step-by-step. The simulator begins first by reading all of the configuration files that the user created for the particular simulation. The configuration files are simple text-based files that list parameters or features and assign a value to them. The value can be a simple number, a number followed by a floating point number that represents the compound annual growth rate for the number, or a function that expresses a more complex operation. Complex operations can include lambda functions that allow the user to express complex changes over the duration of the simulation. A complex operation may also be a value called a *Range* value. We use Range values to describe parameters over which the simulator should permute and repeat multiple operations of the simulator, each with different values from the permutation. Once the simulator reads the configuration files, it checks for any Range values that call for permutation, and the simulator enqueues the permutations so that they can be run in parallel with each other as distinct simulations.

Once the simulator has a list of permutations with which to run each distinct simulation, the simulator spawns new copies of itself to run each simulation separately. We run simulations in parallel in order to take advantage of the parallelism offered within modern computers. As each new simulation starts, if reads its own configuration file that has been created for it by the simulation instance that spawned it. The ability to permute over different values defined in the configuration allows the user of the simulator to explore a large number of possible configurations for the simulations. For example, we used the Range values to explore many possibilities of RAID parity and workload demands. We can also assign Range values to most parameters in the configuration files, but we keep in mind that the number of simulations grows geometrically with the number of parameters assigned to Range values. Each simulation starts, reads its own configuration file, and proceeds to build its simulated archival storage system.

Each archival storage system contains a group of devices. The simulator first calculates how much storage capacity and throughput it will need at the start of the simulation. With these numbers, the simulator calculates how many storage media are needed to meet the capacity demand and how many storage drives are needed to meet the throughput demand of the archival system. If the media and drives are on the same device, we use the larger of the capacity and throughput numbers to proceed to the next step of execution.

The simulator next attempts to install the storage drives and media into the archival system. As it does so, it checks the configuration parameters for each device to find a list of what other devices are needed to utilize the device or media. Devices and media both require libraries, network attached storage systems, or robots. The simulator gathers the necessary hardware and adds those devices to the archival system. Once the devices have been added to the archival system, one new Event is created for each new device. We describe Events in greater detail in Chapter 3.2.1. The simulator adds each new event to a priority queue, and the priority queue ensures that each new device is installed within the simulated archival storage system in the correct order. Tape media can be installed at the same time as tape drives, for instance, but both media and drives must be installed after robotic systems. Each event in the simulator begins and ends at a specific time in the simulation, and the simulator keeps track of time so that it can wake up, resume, and complete Events that may take a long time to complete.

As events in the simulator become completed, the simulator proceeds through time by incrementing its internal clock by the simulation's time slice or epoch. As each new epoch begins, the simulator checks if any new workload Events should be triggered. Workload Events are similar to those when new devices are added to the archive, but the workload events require multiple components of infrastructure to be in place before it can proceed. If the simulator reads data, the drives involved in the operation must be first installed within the archive and also available to use. The simulator also checks for any failed devices inside the simulation. If there are failed devices, the simulator removes the old failed devices, installs new devices by creating an Event to install them, and finally, the simulator creates workload events to simulate writing data onto the new devices for the first time. Once the simulator has done everything it can at the current epoch, the simulation continues by incrementing the current epoch.

When the simulator increments the epoch, it is simulating the passage of time and keeping track of everything that happens in the simulation. When the simulation reaches the end of the time it was meant to run, the simulator creates a data file that contains all of the cost, performance, and behavioral data from the simulation. The data file contains the cost of the archival system at each epoch during the simulation, the total cost of the archive, the amount of power used, the number and types of each device in the simulation, and numerous other facts. Finally the simulator saves the data file with the same prefix as the configuration file that was used to start it. Doing so allows us to return to the data files and determine what configuration parameters were used to produce the data file.

Due to the complexity of the simulation, each execution of the simulation can take an hour or more, depending on how many events, devices, and workload events must be performed during the simulation.

3.1.3 Compound Annual Growth Rates

Many of the factors that drive the demand for and supply of storage in archival systems change over time. Factors such as the capacity and throughput demands of an archival system, the capacity and features of storage devices, and the probability of failure for individual devices as they age may change with time, and furthermore their rates of change or growth may also vary as a function of time. We examine the compound annual growth rate (*CAGR*) of each parameter to describe the rate at which it changes over time based on historical data. We calculate the CAGR for each parameter using the following formula:

$$\text{CAGR} = \left(\frac{\text{val}_i}{\text{val}_0}\right)^{\frac{1}{t_i - t_0}} - 1, \tag{3.3}$$

where val_0 and t_0 are the parameter's initial value and the starting time in years, and val_i and t_i are the parameter's final value and time in years, respectively. We use the CAGR formula to determine growth rates for parameters used in simulations since it calculates the annual growth rate needed to reach val_i, starting with val_0, after $t_i - t_0$ years.

CAGR has been used previously to describe changes in storage technology. Kryder's Law, which describes the growth of hard disk areal bit density, predicts that HDD capacity will double every 18 months, a CAGR of 58% [133]. Hard disk capacity has increased increased more slowly than the prediction of Kryder's Law in recent years [45, 110], so we rely on our observations from historical developments to derive the CAGR for HDD capacity and other parameters with the formula:

$$\text{val}_i = \text{val}_0 \times (\text{CAGR} + 1)^{t_i - t_0}, \tag{3.4}$$

where val_0 and t_0 are the starting value and time, and val_i is the value after a growth rate of CAGR + 1 for $t_i - t_0$ years.

3.2 Simulator Design

Our simulator consists of several modules that we configure to model the functionality of archival systems. Here we present a summary of each critical component.

3.2.1 Events and Event Driver

We use an event-driven model to simulate actions that occur at specified time intervals or with a certain probability over time. Each action is represented by an Event object. Events include installing devices, reading and writing data, and replacing devices that have failed. Each Event acts upon the devices within the simulated archival system, and the status of the device changes accordingly to ensure that no conflicting actions utilize the same resource at the same time.

3.2.2 Time

Time within the simulator affects each of the Events that act upon the archival system. Time progresses in the simulation when, for any particular epoch in the simulation, there are no Events that can execute because either all Events have been completed and closed, all of the Events have a completion time that is later than the current simulation epoch, or if the queued Events must wait for a resource like a drive or storage media to become available. Time progresses in the simulator by incrementing the current time by the length of an epoch. The passage of time in the simulator triggers new Events and adds them to a queue. The current time in the simulator also controls the expiration of Events and values that change with time. Events to install devices, for instance, require a certain amount of time to be completed before the affected device can be utilized for another event. The simulation time affects values such as capacity, failure probability, and device features that change as a function of time.

3.2.3 Archival System and Devices

The Archival System class coordinates actions within the simulator and forwards actions to each individual storage device. The Archival System provides functions to calculate the total capacity, read and write throughput, reliability, and cost of the archival system. The Archival System class also calculates the total cost of meeting a threshold requirement of capacity and performance.

Devices within the simulator include drives and media, networking infrastructure, robots, and racks. We implement a class for each device type to model its unique behavior and features. Devices with removable media such as tape and optical disc use separate Device classes to represent the media and drives. Devices such as hard disk and solid state disk, which do not have separable media, are represented as one Device class. Glass and DNA storage, which feature separable media as well as separate devices to read and write data, are represented with classes for media, a drive for reading, and a drive for writing.

3.2.4 Configuration and Parameters

We use two types of configuration parameters to control the behavior of the simulator: archival parameters and device parameters. Archival parameters define the required capacity of the archive and its rate of growth, the performance and workload demand on the archive, the cost of electricity, and the number of years to simulate. The device parameters include the cost of each device, its performance and capacity as functions of time, and the device's probability of failure.

3.3 Archival Parameters

Digital information has grown in scope as the performance of computer systems has increased over time. The demand for archival storage follows the larger trends of ever-increasing amounts of data. We configure our simulation model to start with an initial capacity of 1 PB. Each simulation that we run begins with this same amount of data, and we extend the capacity each year to simulate the constant growth of data in the simulator. Although we can never be certain exactly how quickly the demand for archival capacity will grow, we utilize a CAGR of 30% that has been suggested elsewhere in predictions about the long-term growth of archival data [20]. Finally, we utilize a nominal workload for our simulations that relate to other aspects of archival systems. Each of our simulations models a workload of 10,000 read operations annually. The small number of read operations reflects an archival system that is only infrequently accessed for data retrieval; however, we vary the number of read operations in our experiments in Chapter 5.

3.3.1 The Cost of Electricity

The cost of electricity can be an important consideration for the location of data centers and the storage systems within them. We derive the cost of electricity for our simulation model from the average cost of electricity for commercial customers in the United States. We take data from the Energy Information Administration's monthly report on the cost of electricity to customers [38], and we utilize the report from January in each year. We show in Table 3.2 the average cost of electricity throughout the US, and we calculate for each year from 2007 to 2020 the compound

Table 3.2: Cost of Commercial Electricity in the US [38]

Year	Price per kWh	CAGR to 2021
2021	$0.1133	–
2020	$0.1059	7.0%
2019	$0.1057	3.5%
2018	$0.1056	2.4%
2017	$0.1055	1.8%
2016	$0.1032	1.9%
2015	$0.1030	1.6%
2014	$0.1052	1.1%
2013	$0.1012	1.4%
2012	$0.0989	1.5%
2011	$0.0998	1.3%
2010	$0.0999	1.2%
2009	$0.1026	0.8%
2008	$0.1050	0.6%
2007	$0.0979	1.0%

annual growth rate of electricity cost between that year and 2021. Table 3.2 shows that the cost of electricity has grown more quickly during the past two or three years than it has since 2007; however, the overall trend has be a slow and consistent rise of a little more than 1% annually since 2007. We utilize the cost of electricity from 2020 and the CAGR growth rate of 1% in our simulation model to calculate the cost of electricity.

3.4 Device Parameters

We utilize parameters for the simulation model that describe the capacity, performance, reliability, growth rates, and operation of each storage technology. The device parameters control how each device performs within the simulator and how each device changes over time. Table 3.3 shows parameters for initial capacity in the simulator, the capacity of the device at the end of its planned roadmap, and

Table 3.3: Storage Media Capacity and Reliability

Medium	Capacity	Roadmap	Annual Failure Rate
Tape [3,95]	12 TB	192 TB	0.000075-0.3%
HDD [11]	10 TB	100 TB	1%
SSD [111]	4 TB	100 TB	0.58%
ODD [124]	0.3 TB	1 TB	0.000075%
Glass [8]	100 TB	360 TB	0.01%
DNA [126]	1 TB	–	1%

reliability for each type of storage device.

We assume that the capacity for each storage technology will develop at a rate consistent with its historical record, and we also assume that capacity growth will slow once each technology reaches the end of its development roadmap as published by its manufacturers. The roadmap for each storage technology offers guidance to the storage industry about what the device's manufacturers expect to be possible for any given storage technology, given their current knowledge and understanding about the practical limits for future developments [8,78,86,97,123,126, 128].

We list the baseline values for each storage technology in Table 3.3 together with the roadmap capacity for each technology after which subsequent increases in capacity grow smaller due to the technical difficulties of further development, which we described in Chapter 2.2. Parameters for glass and DNA storage are estimations based on published research for each technology.

Figure 3.1 shows the expected growth of each storage technology's future capacity that we use in our simulations; values are normalized to the baseline capacity for each technology as given in Table 3.3. The baseline values for existing technologies in Table 3.3 reflect the state of existing technologies in the year 2020. The baseline of archival glass approximates the features proposed for its design from published research [8]. The baseline values for synthetic DNA, however, approxi-

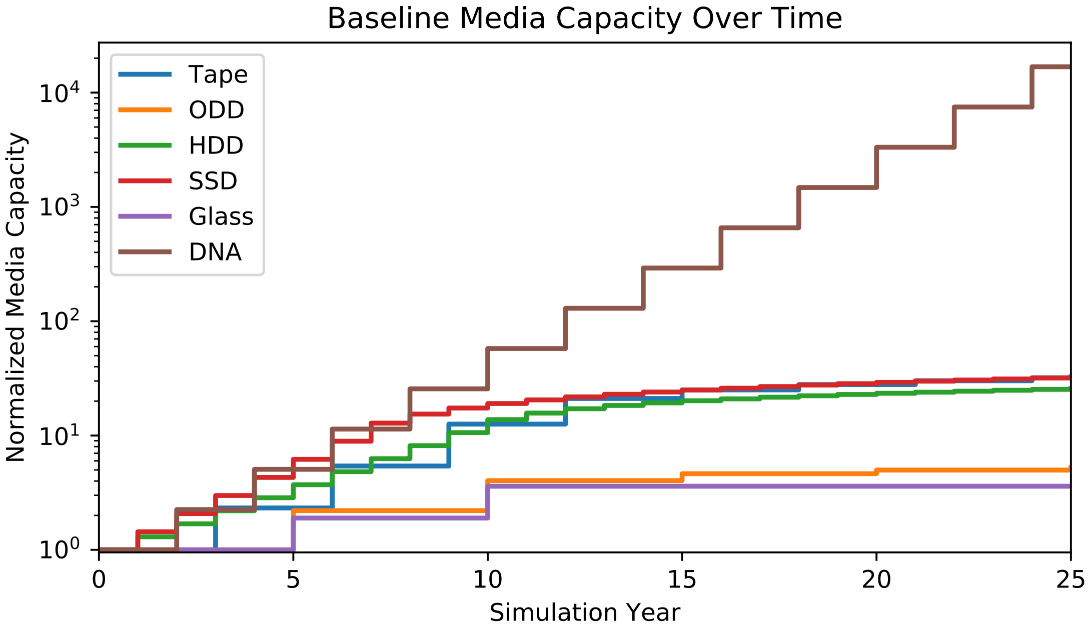

Figure 3.1: We model the growth of device capacity as a step function of time. We present here the baseline growth trajectories for each technology, normalized to the starting capacity for each medium. The rates of capacity growth slow once each technology reaches the end of its developmental roadmap.

Table 3.4: Parameters for Baseline Storage Device Cost

Type	Media	Drive (R/W)	Library	Enclosure
Tape	$150	$8,000	$7,000	$1,000
HDD	–	$200	$10,000	$2,500
SSD	–	$500	$10,000	$2,500
Disc	$10	$10,000	$15,000	$1,000
Glass (est)	$1	$1,000 (r) / $10,000 (w)	$1,000	$1,000
DNA (est)	$100	$1,000 (r) / $9,000 (w)	$1,000	–

mates values for capacity that could make it competitive with other storage technologies, but if DNA should fail to achieve similar performance values in the future, we expect that synthetic DNA could not serve in an archival storage capacity under the constraints that we have envisioned. Nevertheless, synthetic DNA has been developing in terms of both sequencing and synthesis at rates that exceed that of Moore's Law over the past decade [143].

Hard disks and solid state disks frequently increase their capacities as new models incorporate developments and increases in areal bit density for their storage media. We model their capacities in our simulator with annual increases. Tape and optical disc generally offer more infrequent upgrades in part because new generations of storage media require new and often expensive drives. Tape and optical disc have 3 and 5-year upgrade cycles in our simulations, respectively. We expect that archival glass will provide an upgrade trajectory similar to that of optical disc since both optical disc and glass store information optically within a nonvolatile medium. We model the capacity growth of DNA without the constraints of other storage technologies that are constrained by their manufacturing and scalability limitations. Archival DNA's capacity may grow with the developments of the technologies used to sequence and synthesize DNA molecules, and therefore, we model a 2-year upgrade cycle for DNA without tapering the pace at which its capacity grows.

Synthetic DNA has been proposed for archival data storage primarily for the capacity that it could potentially deliver. The performance of DNA-based storage remains today far lower than needed for any realistic storage system [139]; however, DNA is estimated to have a theoretical maximum capacity of 215 PB per gram [55], which is far higher than any other technology that we have considered. Attempts to utilize part of its large theoretical capacity in practical storage systems have so far yielded incremental developments [30, 34, 97]. Recent efforts to develop DNA as a data storage technology have proposed schemes to translate between binary and nucleic acid encoding mechanisms [139]. Other work has proposed methods for isolating and identifying specific DNA molecules that store data in order to facilitate the addressing and retrieval of data that are commonplace with other storage technologies [34, 97]. The capacity of DNA reported in DNA storage research that implement an encoding scheme and addressing and retrieval mechanisms range widely from 83 kB [34] to 1 TB [97]. In any case, DNA storage technology must make large improvements before it can compete with other storage technologies.

Synthetic DNA has been developing at a rapid pace over recent years thanks in large part to the efforts of the medical industry [139]. The cost of DNA sequencing and synthesizing have both declined over time [139], but the current cost of both sequencing and synthesizing strongly disfavors DNA for today's archival storage systems. As recently as 2017, the cost of synthesizing a single base pair for DNA-based storage was estimated at approximately \$0.0001, with generous assumptions for input costs and the availability of synthesis equipment [28]. If we take a base pair to encode a single bit, then the cost per gigabyte of data storage would be $\$0.0001 \times 8 \times 10^9 = \$800,000$ per gigabyte. SSD is the most expensive existing storage technology that we consider in terms of cost-per-byte, and we can therefore compare DNA with SSD in order to underscore the position of DNA relative to every other storage technology. If we assume, as we do in our simulations, that SSDs offer a capacity of 4 TB at a cost of \$500, then SSDs have a cost-per-gigabyte of $\$500 \div 4000 = \$0.125$ per gigabyte. Comparing the two costs, we see that DNA is $800,000 \div 0.125 = 6.4 \times 10^6$ times more expensive than SSD. Of course this difference does not take into account the dramatically slower performance that DNA delivers or the cost of machines to sequence DNA data. Thus, the difference of 6.4×10^6 is a lower bound on the cost disparity between DNA and the most expensive of all existing storage technologies. In order for DNA to compete with other technologies, it will need to improve by at least this much, plus any additional improvements that competing storage technologies deliver over the time that it takes DNA to close the cost and performance gap.

Since we are most interested in considering how and not when DNA could be used for archival storage, we assume for the purposes of our experiments that DNA overcomes its large performance and cost deficits to deliver 100 MB per second of synthesis throughput, 300 MB per second of sequencing throughput, and a latency of one hour. Each throughput value increases by 50% annually in our simulations, a high rate of development that we base on the history of rapid improvements in DNA technology [55, 139]. We also utilize an initial DNA capacity per DNA media of 1 TB with a 50% CAGR. We further assume that the combined cost of synthesizer and sequencer machines will be no more than $10,000, as stated in Microsoft's demonstration of an end-to-end storage system using DNA [126]. Our results for DNA must be regarded in the light of the constraints that the technology imposes on its use: namely, its high latency, separate sequencer and synthesizer machines, and potential for high capacity.

Beyond glass and synthetic DNA, we have envisioned that existing storage technologies will compete with each other for share in the archival storage market. Tape, hard disk, solid state disk, and optical disc each deliver their own features and compromises. Until now, we have not mentioned storage class memories (*SCM*) insofar as candidates for archival storage systems. SCM relates closely with NAND flash, since both technologies are solid state in their design because they have no moving parts. 3D Xpoint [46, 63, 140, 142] in one such SCM technology that delivers performance and endurance beyond that of NAND flash, but the challenge with

SCM technologies for archival storage systems is not their performance. Rather, SCM technologies are more expensive than even NAND flash [46, 63, 140, 142]. We therefore treat SCM devices alongside NAND flash within the SSD category of devices. Nevertheless, should SCM-based devices become viable for archival storage in the future, their cost and performance could deliver the first universal memory that would be equally suited for CPU caching or system memory as long-term archival storage.

3.5 Confidence Intervals

The archival simulator that we have designed for our experiments utilizes pseudorandom numbers to simulate device failures and workload activity. For this reason, each separate execution of the simulator produces unique results. In order to measure the confidence that we have in the results of our simulator, we executed the simulator 100 times using each storage technology and our baseline parameters. Next, we select a subset of those results ranging from three unique simulations to 100 simulations to calculate the size of the confidence interval with as a function of the number of repeated simulations. Finally, we divide the confidence interval by the average total cost of ownership for the archival system after 25 years of operation to determine the size of the confidence interval relative to the cost of each storage technology. We present our results in Figure 3.2.

Figure 3.2 shows that each storage technology follows its own pattern for the size of its confidence interval relative to its total cost. For any storage technology, repeating the simulation 10 times approaches a stable confidence interval. Tape and archival glass present the highest ratio of confidence interval to the total cost of ownership because, for both tape and glass, the reliability of the storage technology is such that drive failures happen only rarely. Nevertheless, the cost of a

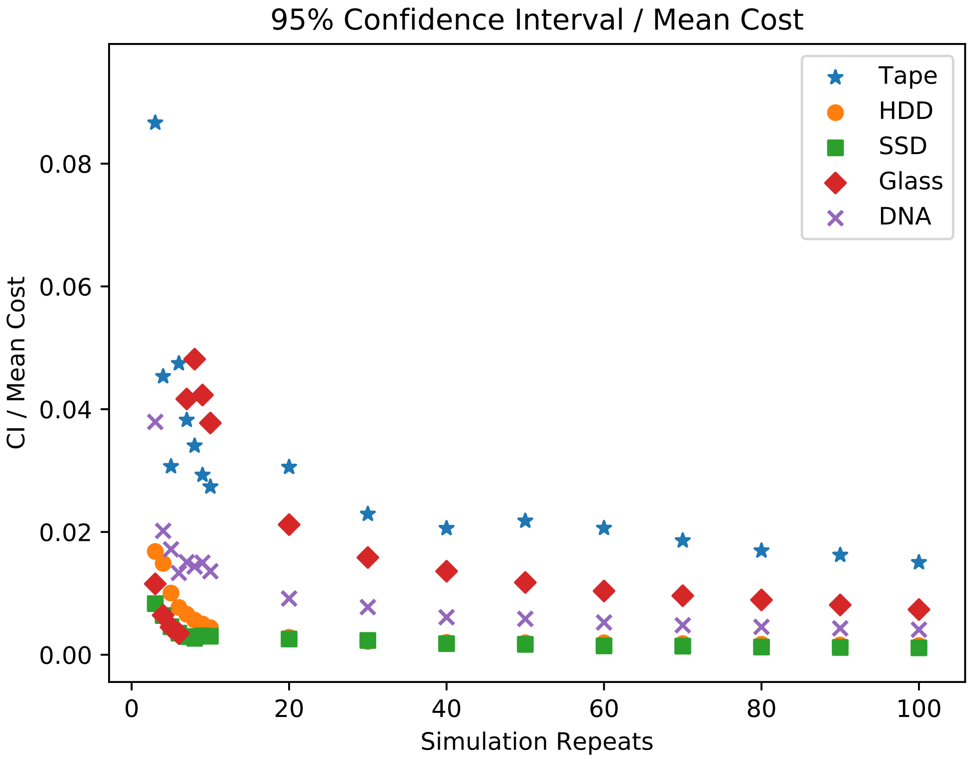

Figure 3.2: The ratio of 95% confidence interval to average cost decreases as we increase the number of separate simulations for a given set of parameters. Tape and glass present the highest ratio of confidence interval to their total cost due to the cost and infrequency of each failure event. Rare and expensive events can have a noticeable effect on the size of the confidence interval within our simulations. Glass sometimes has a low confidence interval for small numbers of simulation runs because glass drive failures occur with such infrequency that they may not happen during each simulation, and the confidence interval with such simulations is smaller than for groups of repeated simulations that capture at least one glass drive failure among them.

failed tape drive or glass write drive is much higher than the cost of a single hard disk or SSD. Thus, each failure event for tape and glass drives causes more inflation to the total cost of the archival system, and because such failures are rare, they do not occur in the same numbers during each simulation. We therefore see more variance between separate simulations using tape and glass than we do for other technologies, particularly those with combined media and drive mechanisms.

Most of the figures in the following chapters do not show confidence intervals. We omit showing confidence intervals in order to make each figure more readable. Nevertheless, we expect that the values for each simulation to fall within 5% of what is shown for each experiment.

Chapter 4

The Cost of Reliability

4.1 Overview

The need for reliable data storage often guides the design and deployment of archival storage systems. As archival systems grow to store an ever-increasing amount of data, so does the importance of reliability as an essential design feature. The risk of losing data within an archival storage system increases with the size of the system and the length of time over which it operates. For this reason, numerous techniques exist to reduce the probability of data loss.

Replication, RAID, erasure coding, and the use of intrinsically reliable storage technologies can serve to increase the reliability of a storage system overall. Archival systems, which typically serve low-intensity workloads over a long period of time, must also integrate available storage technologies and organize them into a design that offers the needed storage capacity and reliability while minimizing the acquisition and operating costs of the archive. Balancing such requirements presents a significant challenge for archival system designers due to the myriad options available for archival storage technologies and configurations, changes to device performance and reliability over time, and the different possibilities for device development in the future. Seemingly simple choices between different storage technologies to use in an archive inevitably yield complex and possibly unforeseen tradeoffs that may result in unfavorable outcomes in terms of insufficient storage reliability or money wasted while achieving a given threshold for archival reliability. Furthermore, predicting the long-term cost of an archive that meets its designers' needs for reliability further complicates the design and implementation of archival systems.

In this chapter, we simulate and measure the cost of acquiring and operating an archival storage system, comparing the cost of using different storage technologies to achieve various levels of storage reliability over time. We include existing storage technologies—tape, optical disc (ODD), hard disk (HDD), and solid state disk (SSD)—as well as prospective storage technologies that are currently under development for archival storage systems—glass and synthetic DNA. We also slightly modify the simulator to support our experiments on archival reliability.

4.2 Approach

Chapter 2.2 described the unique features, limitations, and rates of development for each storage technology within the scope of our simulations, and Chapter 4.3 enumerates parameters of reliability over time for each storage technology. We utilize each device's parameters and enable features within the simulator to measure an archival system's predicted statistical reliability.

The reliability of a storage system depends upon that of the devices in the storage system, the organization of the system, and other events and conditions outside of the system. Events and conditions outside the storage system include software errors, user errors, natural disasters that affect a data center, and electrical faults or surges. While external events are important to storage system reliability [75], their occurrence is not intrinsic to the design or implementation of any storage system since virtually any storage system can be affected by such events. We therefore focus on modeling the reliability of storage devices used in an archival system as well as the design of the archival storage system that can optimize it for reliable long-term operation.

In an archival storage system with many storage devices, each device presents its own probability of failure that typically changes over its lifetime [3, 6, 10, 11, 13, 95]. We define a device failure to be the condition of a device that either cannot reliably write or read data or that has diminished performance relative to its manufacturer's specification. The failure of a storage media device may—but not necessarily—lead to data loss, particularly if the device fails with a degradation of its performance shortly before it ceases to function completely. For the purposes of this work, we compare the reliability of storage devices within RAID groups in terms of their probability of failure resulting in data loss and in terms of the amount of data lost with a failure.

RAID-based storage systems may be configured to replicate data across some number of devices to decrease the probability that any combination of device failures will result in data loss [104]. Various RAID levels combine Reed-Solomon codes with data distribution across multiple devices to offer configurable levels of performance and data reliability [104]. The probability of data loss in RAID systems depends on the probability that multiple devices within a RAID group will fail during the time required to recover from a failure [71–74, 102, 104]. RAID levels with extra redundancy or parity can tolerate more near-simultaneous device failures, but they require more storage devices to store a given amount of data than RAID levels with less redundancy or parity. Archival system designers balance performance and reliability versus cost to suit their particular needs.

Patterson, et al. derived a method for calculating the reliability of a storage system based on RAID parity [104]. The Patterson model finds the mean time to failure for RAID groups based on the uniform mean time to failure of devices similar to those in each group [104], thereby assuming that each device presents an identical probability of failure to every other at any given moment in time. We observe in Chapter 4.3 that not all devices that operate within any storage system present the same probability of failure at all times. In particular, hard disk drives—the storage technology for which RAID was originally designed—prove more reliable at some times during their lifetimes than at others. Hard disk failure rate has been described as a *bathtub curve* to represent the higher probability of failure at early and late stages of its life [33]. Storage devices fail with probabilities that are unique and independent for each device, and we calculate the overall reliability of each storage system based on the unique life cycle of each individual storage device. Our goal for evaluating the reliability of a storage system must be to first establish the probability that one or more device failures will result in the loss of data.

We calculate the probability that two independent events occur together with the formula:

$$P(A \cap B) = P(A) \times P(B), \tag{4.1}$$

where A and B are independent events, and $P(A)$ and $P(B)$ are the probabilities of each event occurring on its own [67, 134]. We can generalize Formula 4.1 for an arbitrary number of independent events by adding more independent variables and multiplying them with the others. As we will observe in Chapter 4.3, storage device reliability changes over time as each device ages. Therefore we consider each device failure as its own independent event, allowing us to deploy Formula 4.1 in our calculations of storage system reliability. Although batch failures can increase the probability of data loss in practice [101, 102], we limit the scope of our model to rare and uncorrelated failures.

The reliability for a storage system with k parity devices has been derived and provided in other work as:

$$R(T_k) = 1 - \prod_{i=1}^{k} \left(1 - e^{-\lambda_i \times T_k}\right), \tag{4.2}$$

where λ is the failure rate for device i and T_k is a time interval over which to evaluate the reliability of the storage system [71–74, 102]. Formula 4.2 is a generalization whereby many probabilities can be combined using Formula 4.1 to generate a composite overall reliability.

Duritg the simulations relating to reliability, we utilize a simplified model for calculating the failure probability for each device, and we utilize a simplified calculation in order to expedite the complex operation of our simulation model. When considering the reliability of each device, our simulation model evaluates each device individually at each time interval in the simulation. Our simulation model supports time intervals ranging from one second up to one year, and our simplified approximation of Formula 4.2 helps to improve the performance of the simulation model while preserving a close approximation for the calculation in Formula 4.2.

For each device in the simulation model, we calculate its failure status using the inequality:

$$\text{Random} < \lambda \times t, \tag{4.3}$$

where Random is a pseudorandom number in the range from 0 to 1, λ is the annual failure rate (*AFR*) for the device, and t is the length of each simulation epoch. For smaller values of t, the probability of failure during each epoch decreases; however, shorter epochs also ocurr in greater numbers within the simulator. We present in Figure 4.1 a comparison of the standard failure model as given in Formula 4.2 with

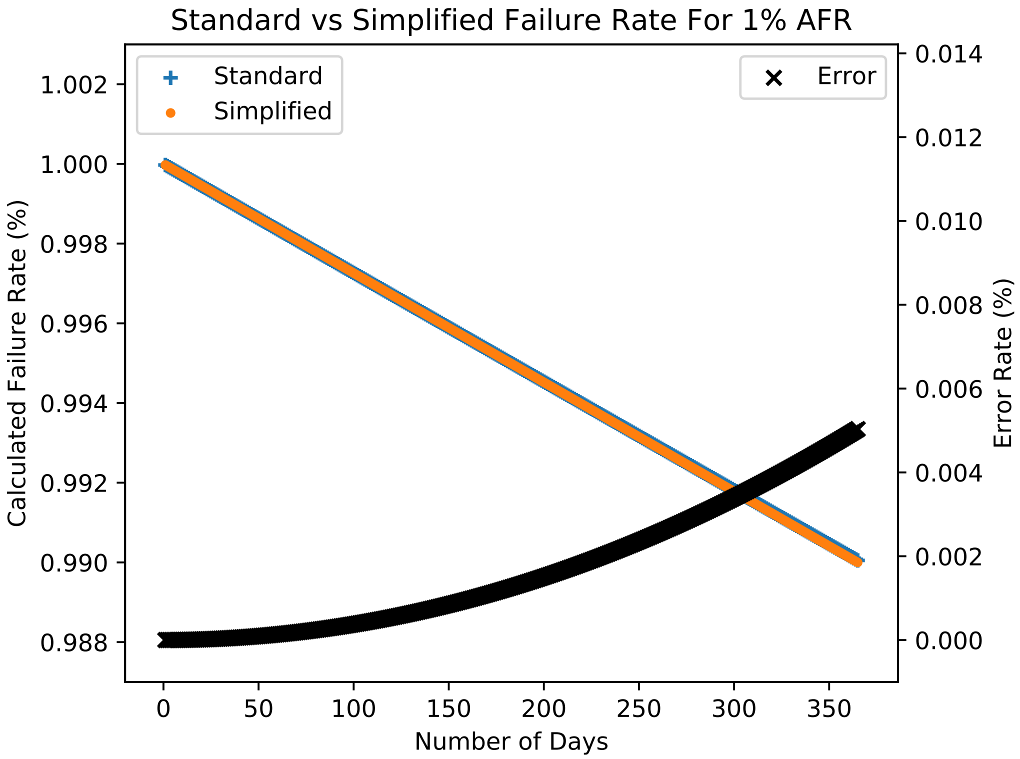

Figure 4.1: We compare the standard reliability model with our simplified approximation. We show the absolute AFR values as calculated using each technique in the y-axis on the left. On the right we show the percent error that separates our approximation from the standard model.

the simplified model given in Formula 4.3. In this example, we use a device annual failure rate of 1%, which is similar to the AFR values for the devices that we have modeled with our simulator.

Figure 4.1 demonstrates that our approximation for the failure probability of each device closely follows the standard reliability model as shown in Formula 4.2. Furthermore, as we use an epoch length of one day in our simulations, the error that we would encounter will also be the smallest as shown in the left of Figure 4.1.

4.2.1 Blast Radius

Ideally, the probability of failure for any storage system would decrease as the amount of data that would be lost from a failure increases; however, such a tradeoff proves difficult to achieve with many storage technologies. The potential amount of data lost depends on the amount of data in a RAID group, which itself depends upon how many data drives are used in the group as well as the capacity of those drives.

As the capacity of each storage device increases with time and development, so too does the capacity and rebuild time of each RAID group, assuming a fixed number of drives in each group. Using high capacity storage devices in a small number of large-capacity RAID groups introduces a greater risk for catastrophic data loss than many smaller RAID groups would. Large storage devices and RAID groups necessarily increase the amount of data that would be lost during any failure. By Formula 4.2, a small number of RAID groups may present a lower total probability of data loss due to a RAID group failure than a larger number of RAID groups would; however, reducing the number of RAID groups requires increasing each group's capacity to store a given amount of data. Rebuild times for larger-capacity groups are also longer, which offsets the reliability advantages of reducing the number of groups by using larger drives. Maximizing reliability in a storage system therefore requires balancing the use of many small RAID groups with fast rebuild times and larger RAID groups with lengthy rebuild times. Formula 4.2 also disguises the amount of data that would be lost as a result of any single failure, however rarely that may happen.

Storage systems and RAID groups *should* become more reliable as they store more data. Existing storage technologies continue to develop apace, yet the reliability of each device has not increased as quickly as its capacity. Prospective storage technologies like glass and synthetic DNA also promise large capacities, and their exact reliability remains uncertain. High capacity storage technologies allow increasing amounts of data to be concentrated within each RAID group; however, as the total probability of failure $R(T_k)$ decreases with fewer RAID groups that each have greater capacity, the probability of failure relative to capacity may increase if capacity grows in each group more quickly than reliability increases. In order to model the relationship between group capacity and reliability, we must utilize a metric that incorporates group capacity alongside reliability. We define the probability of failure relative to capacity with the formula:

$$F(s,t) = P_s \times C_s, \tag{4.4}$$

where $F(s,t)$ is the probability of failure relative to capacity, P_s is the probability of a storage device or RAID group s failing over time t, and C_s is the storage capacity of device or group s. Formula 4.4 quantifies in $F(s,t)$ the relationship between capacity and reliability such that storage groups with relatively low reliability and low capacity can deliver a lower value for $F(s,t)$ than storage groups with much higher capacity. By Formula 4.4, a RAID group that stores 1.000 TB should offer a lower probability of failure p by a factor of at least 100 than a RAID group with a capacity of 10 TB in order to compensate for its larger capacity. The value of $F(s,t)$ conveys a numerical representation of how much risk lingers within the storage system that a failure will result in a catastrophic loss of data. We describe the risk of catastrophic data loss as the *blast radius* of a storage group or a storage system. The blast radius of a RAID group is equal to its failure probability relative to its capacity as given in Formula 4.4.

4.3 Simulation Parameters

We simulate the reliability, performance, and cost of archival storage systems using parameter values that describe candidate archival storage technologies. Our simulator uses configuration parameters to define the performance, capacity, reliability, and cost of storage devices that may be used within a storage system. The output of our simulator therefore depends upon the parameters that we use for each type of archival storage device. In this section, we present details of each storage technology's reliability and prospect for future development.

4.3.1 Archival Tape

Tape is a popular archival storage medium due to its high capacity, reliability, stability when stored for long periods of time, and good performance on sequential workloads. Its weaknesses are its poor random access performance, the high cost of tape drives compared with tape media, and the length of time it can require to access information—the time retrieve to the first byte.

Tape Reliability

One of the main advantages of tape as an archival medium is its ability to cost-effectively store large amounts of data reliably and with minimal need for ongoing maintenance. A 2012 study of the tape archive at the National Energy Research Scientific Computing Center (NERSC), which consisted of 40,000 tape cartridges that were between two and 12 years old, showed a reliability rate of 99.9991% when reading data [3]. NERSC relied upon a single copy of data within its archive, a choice facilitated by tape's high sequential read and write speeds as well as its high reliability as observed in the NERSC archive. The workload on the NERSC archive included a 30% daily read rate, which is much greater than many other archival workloads. While such a workload requires a tape archive to perform as though it were primary storage, it also serves the purpose of quickly discovering any problems that arise in the archive by continuously scrubbing or verifying the archive's data during each read operation, and continuous scrubbing preserves the archive's reliability by verifying that data is readable and not corrupted. We use the NERSC study for our optimistic tape experiments with a failure rate of 0.0009% over 12 years or 0.000075% annually.

Another study of over 1 million tape cartridges shows that nearly 5% have at least one unrecoverable bit error during their lifetimes while 0.3% have at least 10 unrecoverable bit errors [95]. In our experiments, we set the pessimistic tape reliability to have a 0.3% annualized failure rate. The study finds that removing the least reliable 3% of tape cartridges could significantly improve the reliability of the tape-based archive as a whole. Tape as a storage medium is particularly sensitive to the environment in which it is stored and used; work continues on studying the impact of environmental pressures on the reliability of tape and how environmental conditions should inform the design of tape-based archival systems [6].

Tape Development

Tape has been developing consistently since its first introduction as a storage medium. The popular LTO-8 format of tape cartridges features 12 TB of storage capacity and approximately 300 MB per second of read and write throughput [50]. New generations of tape become available every two to three years with each generation of tape drive being able to read tape cartridges that are one or two generations older than it. The LTO Ultrium consortium has published a roadmap for increasing tape cartridge capacity to 192 TB [21] within 10 to 15 years.

4.3.2 Hard Disk

Hard disk drives (*HDDs*) are a popular medium for long-term archival storage due to their high capacity, widespread availability, and adaptability to sequential and random-access workloads. The difficulty of using them within archival systems includes their lower reliability compared with tape and recently their lower rates of development for capacity and performance.

HDD Reliability

We examined the reliability of hard disk drives using the Backblaze hard drive data set [11]. We analyzed six years of the dataset to measure the observed reliability and life cycle of hard drives in an online backup setting. Our goal for analyzing the Backblaze data is to gather insight on how hard disk reliability may be changing over time. We also observe the way in which trends in hard drive developments affect decisions about device retirement and replacement within Backblaze's server infrastructure. We apply these insights to our simulation model.

We began our data analysis by importing the hard drive statistics available from Backblaze. Next, we removed from the dataset of all of the drives that were active within the Backblaze data center on the first day for which data is available: April 10, 2013. We removed all of these drives from our analysis because we do not know from the information available how many drives failed or were removed before that first day in the dataset, and including them would introduce a bias to our observations. We measured all drives by the date that they were added into the Backblaze storage system, and we counted how many days each drive was active in the system until it either failed or was retired for another reason. A drive was said to have failed when the dataset's marker for failure became true. A drive was said to have been retired after the last day it appeared in the dataset and if the drive was

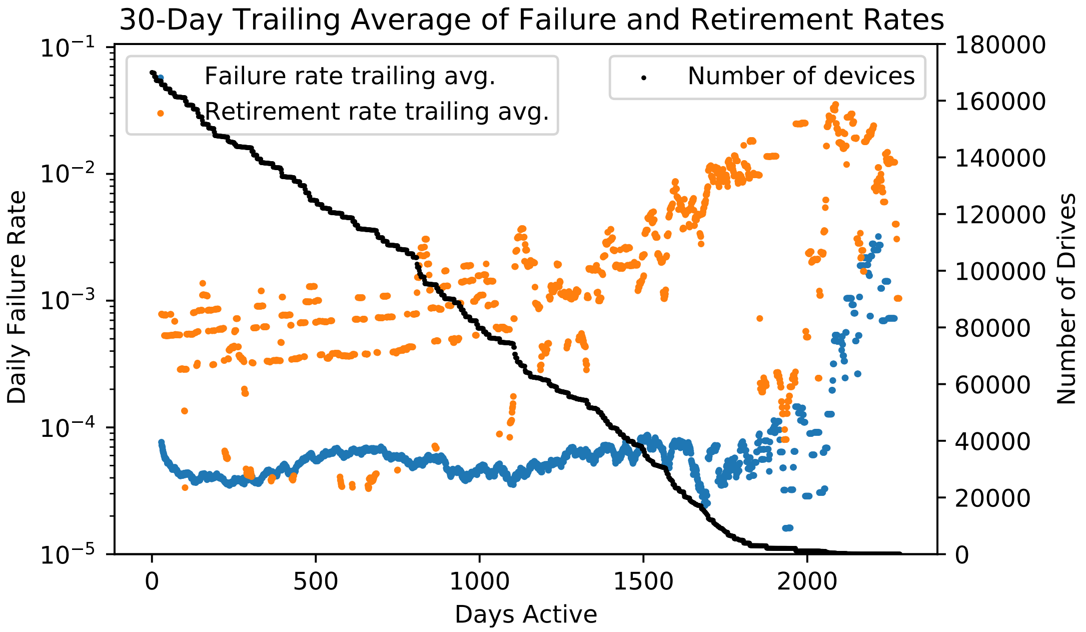

Figure 4.2: The first day on the x-axis begins for each drive when it was added to the Backblaze storage system. Failure and retirement rates are shown in percentages. The number of drives over time shows the number that survived in the data center as drive age increased.

not already marked as failed. Next, we determined how many drives survived for a given number of days. We also calculated how many drives failed or were removed from Backblaze's system after a given number of days. Finally, we determined the daily failure and retirement rates for all hard drives as shown in Figure 4.2. We also show in Figure 4.3 the cumulative portion of hard drives that were active, retired, or failed over their lifetimes and based on the calendar year in which the drive was first added to the storage system.

Figure 4.2 shows the 30-day trailing moving average of the daily hard drive failure and retirement rates. We observe that the failure rate for all drives combined remains stable until after the drives have reached approximately 5 years of age. The failure rate begins to increase after five years as the drives reach the end of their warranty periods. In Figure 4.3, we observe that hard disk failures grew more

Table 4.1: Number of Operational Days Before Reaching HDD Failure Rates

	Drives Added During Year					
%	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
1%	67	143	168	268	332	252
2%	208	332	357	565	589	440
3%	394	503	554	948	745	582
4%	491	684	721	1349	1104	–
5%	618	816	941	–	–	–
10%	1195	–	–	–	–	–
15%	1793	–	–	–	–	–

quickly in 2013 and 2014 than they did in 2017 and 2018. Table 4.1 shows the number of days taken to reach different failure rates and separated by the year in which the drive was added to the Backblaze data center. Drives added in 2013 took slightly over two months to reach a 1% failure rate, but drives added in 2018 took approximately eight months to reach the same 1% failure rate. We attribute this observation to the improving reliability of hard disk drives in recent years. We also observe that failure rates remain mostly consistent with time up to five years for drives added within each calendar year. Nevertheless, retirements rather than failures are the dominant reason for the removal of hard disks from the Backblaze data center.

We considered three possible reasons why a hard drive would be removed without failing. First, the hard drive may be part of a model or batch of drives that are likely to fail in the near future. In order to preempt multiple simultaneous drive failures that could result in data loss, system administrators may choose to replace the faulty drives with a more reliable model; however, as shown in Figure 4.2, we did not observe a high rate of premature drive failures, nor did the pattern of drive retirements follow any sudden increases in drive failures. A second possible reason for drive retirement is the increasing likelihood of failure as each drive ages, particularly after approximately five years of operation as shown in Figure 4.2. Even though Backblaze's storage system may have enough redundancy to survive a high rate of drive failures, system administrators may prefer to control the timing of data migration between new and old drives rather than repairing drive failures as they occur. A third possible reason for early drive retirement is the availability of higher capacity drives which, if used to replace older drives, offer much greater capacity while using the same or less power.

HDD Development

The ever-increasing capacity and performance of hard drives helps to accelerate the replacement cycle of old drives; however, a reduction in the pace of hard drive development will also reduce the benefits of replacing older drives with new ones. Figure 4.3 shows a comparison of the portion of active, retired, and failed drives throughout their lifetimes and separated by the year in which each drive was added to the Backblaze data center. The portion of active, retired, and failed drives varies over time depending upon which year the drive was added to the storage system. Drives that were added in 2013 and 2014 began to be removed rapidly as they reached 1500 and 1200 active days, respectively; however, the cumulative total of drive failures did not increase dramatically during that time. Drives added in 2015, on the other hand, have been retired much more slowly as they age through 1500 active days compared with those added in 2013 and 2014. Hard drives available in 2018 and 2019 do not offer as compelling of a reason to replace drives from 2015 due to the slow pace at which hard drives have developed between 2015 and 2019; furthermore, the improvement in hard drive reliability supports their continued use as they approach five years of age. We conclude that a slowing growth rate of hard disk capacity will lead to fewer hard disk retirements. Instead, the reliabil-

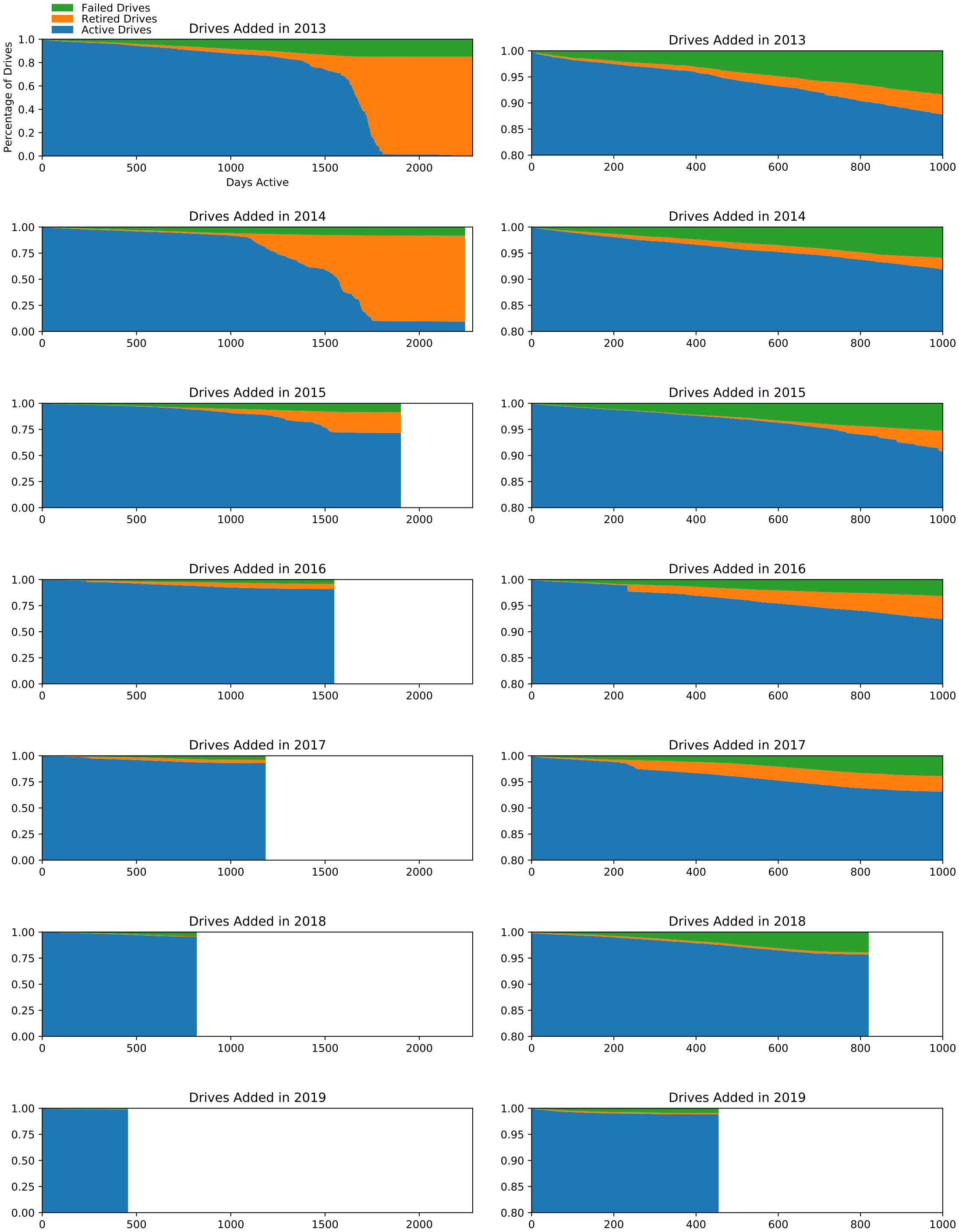

Figure 4.3: Figures on the left show data for all drives that were added to the Backblaze data center within the specified calendar year. Figures on the right show the first 1,000 days and top 20% of the data. The x-axis corresponds to the lifespan for each drive.

ity of hard disks at and beyond five years of age will become increasingly important for disk-based storage systems.

When describing their choices for removing older model hard drives, Backblaze confirmed that 8 TB drives replaced 2 TB drives [65] and 12 TB drives replaced 3 TB drives [66] that were more common in 2013 and 2014 than in 2015. The availability of higher capacity drives and the need for more data capacity within the same data center motivate decisions to upgrade hard drives from older, lower capacity models to newer, higher capacity models. For this reason, the pace of hard disk drive development will, to a large extent, determine the demand for drives within backups and other cold storage systems such as archives.

The International Disk Drive Equipment and Materials Association published a roadmap in 2016 indicating that hard disk drive capacity will continue to scale for years to come. Capacity will increase as manufacturing techniques improve and as existing technologies for increasing density mature [128]. New technologies like Heat-Assisted Magnetic Recording (HAMR) and Heated-Dot Magnetic Recording (HDMR) will improve HDD capacity when they become available; however, since new technologies can be challenging to manufacture reliably at scale, their emergence and the capacity they promise may prove to be uneven and prolonged. As we observe in the Backblaze dataset, significant increases in hard drive capacity motivate the adoption of newer drives. If the availability of higher capacity hard drives becomes increasingly terraced in years to come, we can expect the adoption of new generations of hard drives also to follow an increasingly uneven pattern following the availability of new technologies that increase hard drive capacity.

4.3.3 Solid State Disk

SSDs are ideal for demanding workloads due to their low latency and high throughput. SSDs have become prominent for primary data storage where performance is critical and where cost per gigabyte is not a primary concern. SSDs have also benefited from advances in manufacturing and design that lead to ever greater capacity, improving performance, and excellent reliability.

SSD Reliability

The reliability of SSDs has been studied within the context of demanding data centers. Meza et al. show that SSD failure rate increases non-monotonically with time and with the amount of data written to the device due to multiple different failure modes that dominate at different times during the lifetime of an SSD [88]. Reliability also varies widely with SSD model and the workload on the drive. Schroeder et al. found that the rate of unrecoverable errors grows linearly with the number of program-erase (PE) cycles across multiple SSD models. Furthermore, newer SSD drives offer similar or better reliability compared with older SSD models notwithstanding the smaller lithographies and additional bits per cell in newer drives [113]. SSDs can trade PE endurance for capacity by increasing the number of bits stored per cell of flash. We assume in our simulations that SSDs have an annual failure rate of 0.58% based on figures published in SSD data sheets [111].

SSD Development

SSDs are increasing in capacity over time as their development continues. Over time, declines in the cost of manufacturing each byte of storage dominate the total cost of data storage. SSDs have increased in capacity due to smaller lithographies, by stacking multiple layers of flash cells to form a three-dimensional flash chip, and by increasing the number of bits stored in each cell of flash. Recent additions to the number of bits per cell [81] and the number of layers in each flash chip [82] promise to extend the development of flash-based SSDs into the future.

4.3.4 Optical Disc

Optical disc (ODD) is a mature technology that promises durable and scalable archival storage. Optical disc is less sensitive to its long-term storage environment than other archival technologies such as tape [123]. Disc offers a 50 year lifetime for write-once archival storage media [124], and each new generation of optical drives remains compatible with every previous generation of optical media. Future generations of optical disc will triple capacity from 300 GB to 1 TB per disc [107]. Still, the slow rate of development, limited number of vendors, and the shortage of detailed information about long-term cost and reliability present ongoing challenges to optical disc as an archival storage technology. We assume for the purposes of our simulations that optical disc has a reliability equal to that of optimistic tape.

4.3.5 Archival Glass

Glass has been proposed as a novel long-term storage technology for use in data centers and archival systems [8]. Glass-based storage utilizes femtosecond lasers to encode data into a multi-dimensional pattern within a small plate of glass. Glass could offer good potential as a storage technology due to the low cost of manufacturing the storage medium, high data density, and excellent reliability. Ongoing work is refining the technology by increasing density and throughput for both read and write operations. Unlike tape and optical disc that have separate storage media and drives, glass-based storage requires separate media, drives for reading, and drives for writing.

4.3.6 Archival DNA

DNA has been envisioned as a high-capacity medium for long-term archival data storage [126]. DNA promises storage density that is orders of magnitude greater than existing storage technologies [97]; however, there remain many challenges to implementing a functional storage system that uses DNA as the storage medium [30]. Synthetic DNA requires a synthesizer to encode data within DNA molecules, a repository to store the DNA over a long period of time, and a sequencer to read data from the DNA molecules. Takahashi et al. recently demonstrated an end-to-end storage system with an approximate total cost of $10,000 that synthesizes, stores, and retrieves data using DNA molecules [126]. The principal challenges of utilizing DNA for archival storage remain a high latency for read and especially write operations, the difficulty of encoding data into the language of DNA, the reliability of the storage system, and the cost of materials and equipment.

DNA could potentially store up to exabytes of data per mm^3 [30], potentially over thousands of years [34] if the storage system uses enough redundancy and effective protection from contamination and degradation. Unlike other storage technologies, DNA has been used by biologic organisms to store and transmit information throughout history. DNA does not require migration of data from older to new generations of storage media as time passes because the underlying storage medium remains the same, assuming that the encoding scheme for binary information stored in DNA remains accessible into the future. Erasure encoding schemes have been proposed to protect DNA storage from the possibility that the DNA molecules will degrade over time [30,34,139]. Given some assumptions about its performance and cost, DNA can be compared with other storage technologies on the basis of its cost to store data over time while achieving a needed amount of reliability.

We expect that DNA-based storage systems will remain in development for years before they become commercially viable; however, we begin our simulations for DNA at the current year in order to easily compare it with other technologies while demonstrating how DNA's cost changes with the target reliability of the storage system.

4.4 Experimental Results

We design our reliability experiments to measure the total cost of using each archival storage technology in a variety of RAID configurations over 25 years of operation. We vary the number of data drives in each RAID group, the number of parity drives, and the maximum age of the storage devices before they retire from the storage system. The values for data drives and parity drives in the RAID configuration, retirement age, and the type of storage technology remain unchanged during each simulation. Separate simulations test different combinations of values and storage technologies. We quantify the total expected reliability of each storage system by first calculating the system's reliability against failure during one year of operation. We use the reliability value to determine the number of *nines* of reliability for each storage technology and RAID configuration. We calculate the nines of reliability as the number of consecutive significant digits that are equal to nine above a total storage system reliability of 99%.

We use the minimum reliability and maximum blast radius found during each simulation to represent that simulation's RAID configuration, and the cost of each configuration, measured in US dollars, is the cumulative acquisition and energy cost after 25 years. For each storage technology, we plot the minimum cost to reach zero through 16 nines of reliability. Each simulation begins with an initial minimum capacity of 1 PB that grows by 30% each year.

4.4.1 Reliability Cost Inflation

Our baseline experiments as shown in Figure 4.4 were based on the values in Tables 3.3 and 3.4. We ran our simulations to describe the outcome of various configurations of parity, and as we considered various combinations of parity and data drives, we do not always find a configuration that matches each of the points of the figure. We assume that, for any particular data point, more reliability is better than less reliability, and for each quantity of "nites" of reliability for which we do not find a configuration, we search for configurations that deliver more nines of reliability to fill in and represent the points where we found no configuration. Flat lines between different nines of reliability for each technology indicate that more reliable RAID configurations stand in for lower levels of reliability for which we found no configuration.

We also include separate simulations for hard disks with uniform failure rates and exponential failure rates. A uniform failure rate for hard disks is an unchanging annual failure rate (AFR) of 1%. We abbreviate uniform failure rates as "const. fail". Experiments for hard disks with exponentially-growing failure rates have a uniform AFR of 1% until five years of operation within the archival system, and the failure rate doubles each year thereafter. We label such simulations with "exp. fail". As we discussed in Section 4.3.1, experiments for optimistic tape use an AFR of 0.000075% that does not grow over time while our pessimistic experiments for tape use a 0.3% AFR that grows 10% each year.

We observe that optical disc (ODD) costs the most of all storage technologies due to its limited road map of future developments. We study ODD further in Section 4.4.8. SSDs also have a high cost across the entire range of reliability values because of their high cost per byte of storage relative to other technologies. We explore the cost and reliability of SSDs further in Section 4.4.5.

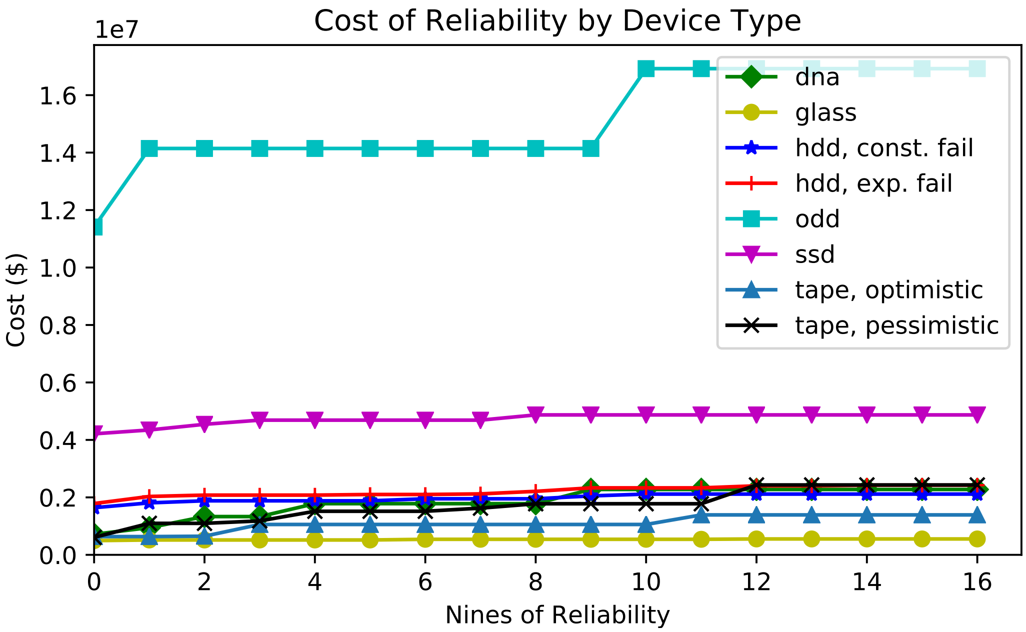

Figure 4.4: The minimum cost of achieving different levels of reliability varies by the type of storage used in an archival system. The cost values are expressed as cumulative total for the archive after 25 years of operation. Most lines look flat on this graph due to the scale set by the high cost of optical disc. The stepped nature of optical disc reflects the significant cost of adding more drives to read and write data over the entire simulation time. Optical disc in particular has high costs for both library systems and drives, and the jumps in cost correspond to the need to add both libraries and drives to spread parity over more devices.

Our pessimistic experiments for tape and hard disk both return similar results for the highest levels of reliability. We conclude from this that hard disk and tape are competitive in terms of cost for reliability in archival storage. The low AFR of optimistic tape requires less RAID parity and therefore fewer tape cartridges to reach 16 nines of reliability than pessimistic tape, and the fewer number of tape cartridges and other hardware like tape drives needed to support them results in a total cost that is 43% lower for our optimistic tape results. If the actual reliability of future tape media is better than our pessimistic AFR of 0.3%, we expect that tape will cost less than hard disk at every level of reliability.

Our experiments for archival glass show that it could become a highly cost-effective storage medium, granted that our assumptions about the cost of glass media and drives prove accurate. We explore other possibilities for glass in Section 4.4.6. Synthetic DNA, on the other hand, struggles to compete with existing technologies due in large part to the high cost of materials for each DNA molecule. We further explore DNA in Section 4.4.7.

We use Formula 4.4 from Section 4.2 to calculate the blast radius for each storage device. We defined the blast radius to be a function of the failure probability of a RAID group relative to its capacity. A large blast radius indicates a high average probability of losing data when a RAID group fails. As shown in Figure 4.5, the blast radius varies widely by storage technology and cost. In this experiment, we present the blast radius for devices with a maximum age of 10 years. Each storage technology has multiple data points since different RAID configurations result in different values for cost, RAID group capacity, and RAID group failure probability.

We observe that archival glass has a relatively large blast radius due to the high capacity of each storage device. For many of the storage technologies like DNA, HDD, and tape, blast radius can be minimized without much additional cost. DNA

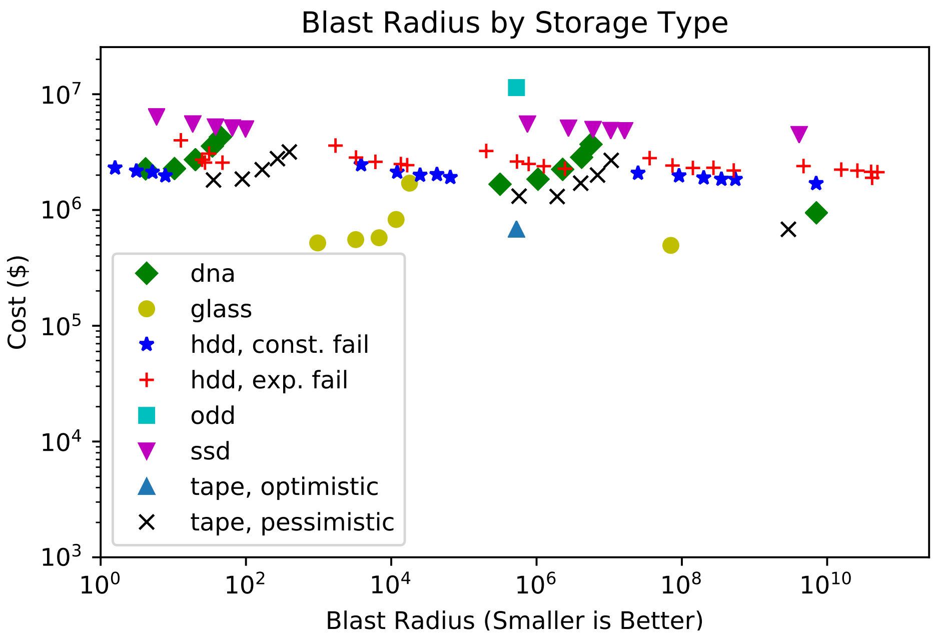

Figure 4.5: The total blast radius for an archival system depends on the reliability of the storage devices, the amount of parity in each RAID group, and the capacity of each device. The lower left quadrant of the figure represents the ideal outcome of both low cost and a low risk of catastrophic data loss. The upper right quadrant of the graph represents the higher costs and higher risk of data loss, which is an undesirable configuration.

and optimistic HDD offer both moderate costs and a low blast radius. Increasing the capacity of either HDD or DNA would necessarily increase the blast radius; however, we propose changes to their designs in Section 4.4.4 and Section 4.4.7 that could further reduce their costs while preserving their low blast radius.

4.4.2 Cost and Reliability of Tape

We showed in Section 4.4.1 that reducing the AFR of tape media can have a large effect on the total cost of an archival system across a range of reliability values. If, however, the AFR of tape storage media increases as its capacity continues to grow with time, how much more will tape archival systems cost while meeting the same reliability goals? We ran simulations with increased AFR values for tape media from 0.5% to 5%. Figure 4.6 shows that the total cost of tape-based archival storage grows with higher AFR values for tape media at each level of reliability; however, a tenfold increase in AFR results in less than a doubling of total cost for 16 nines of reliability. We see an 81% higher total cost with a 5% AFR compared with a 0.5% AFR. We therefore observe that the cost of a highly reliable archive using tape increases more slowly than the AFR of tape media. The large difference between our optimistic and pessimistic experiments for tape in Section 4.4.1 reflect the impact on cost of an increasing AFR as devices age. With all other tape experiments using an AFR that grows 10% annually with device age, the optimistic experiments show that a stable storage medium significantly reduces the cost of highly reliable storage because stable old storage devices can remain in the archive without dramatically increasing the probability of data loss. We conclude that the stability of the AFR for tape can have a large impact on the cost of reliably storing data over the long term.

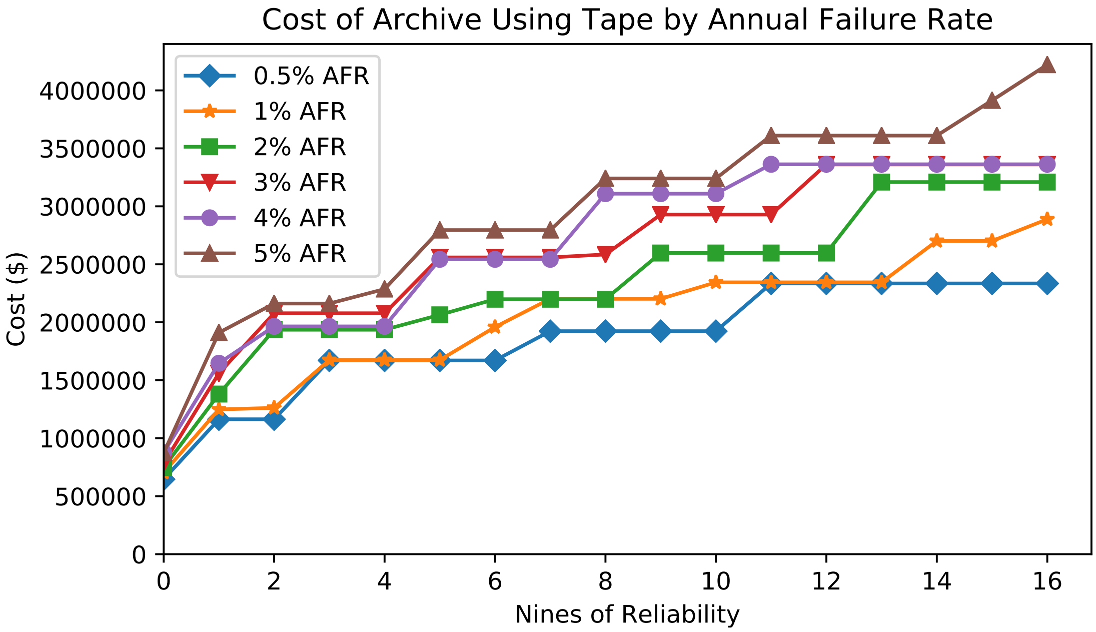

Figure 4.6: The cost of reliability for tape increases marginally with the AFR of tape media.

4.4.3 Hard Disk Reliability

Figure 4.7 shows results of simulations using two models for hard disk reliability. We compare the uniform failure rate with the exponential failure rate for hard disks as described in Section 4.4.1 while also comparing different maximum ages for the drives in the archive. The *max age* is the age at which drives are retired from the storage system and replaced with new drives.

We observe as expected that growing failure rates result in higher costs overall compared with the optimistic case of hard disks that fail with an unchanging AFR of 1%. The most economical option for all levels of reliability with drives that have uniform AFRs is to keep the drives for as many years as possible because such old drives would not fail with any greater probability than new drives. For drives with failure rates that increase after five years of operation, keeping the drives for 11 years proves to be the most expensive option. Instead, it is best to keep drives

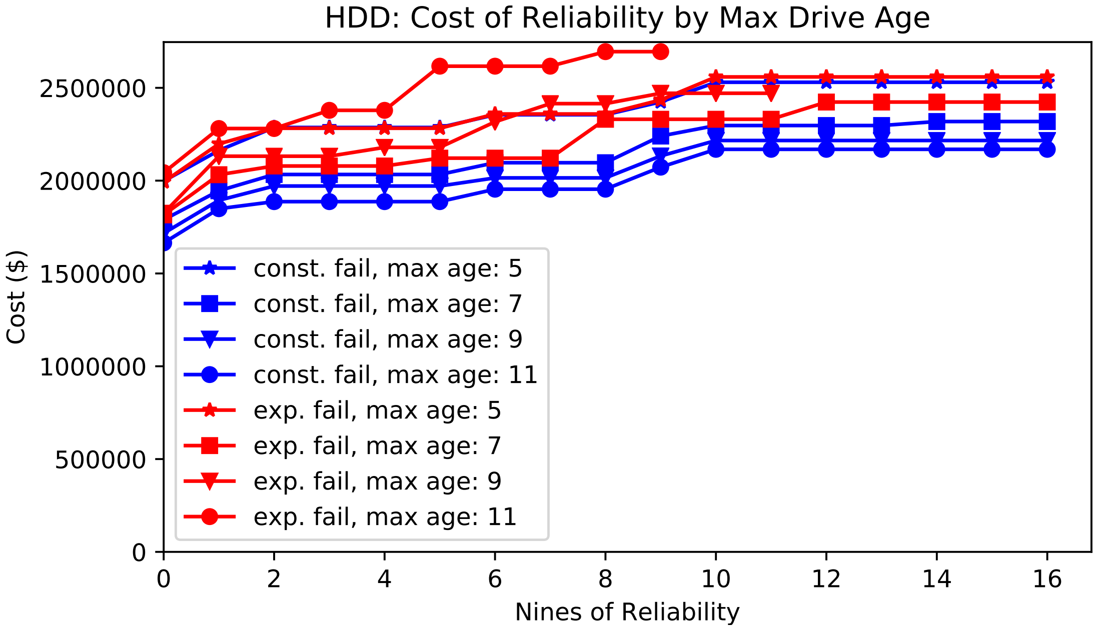

Figure 4.7: Hard disks with exponentially growing failure rates cost more to use than drives with uniform and unchanging failure rates. Lines that do not extend across the entire *x*-axis indicate that we found no RAID configuration to reach those higher levels of reliability.

for approximately seven to nine years in order to minimize the cost of the archive across a range of reliability values while simultaneously extracting as much useful lifetime from each drive as feasible. Keeping the drives for longer reduces the storage system's total reliability so that additional parity must be used to compensate for the increasing failure rate of old drives, and adding more parity drives causes the cost of the system to increase. Uniform failure rates for hard drives reduce the cost of an archive with 16 nines of reliability by 10%. We conclude that if hard disks could be made to last at least 10 years instead of five, the cost of constructing a reliable archive using hard disks would decrease accordingly.

4.4.4 Hard Disks With Removable Media

Hard disk drives currently have physically combined platters and recording devices. We designed experiments to explore the possible benefits to the cost of reliability in archival storage of separating platters from the HDD recording mechanism. In these experiments, we use the same capacity and failure rate as our other experiments with hard disks. We assume that the platters of the drive by themselves will cost 75% of what a typical hard disk costs and that the archive will use similar mounting infrastructure to a tape library system. Finally, we assume that the read and write mechanism for the removable platters will cost more than a traditional hard drive but less than a tape drive. We use the estimate that the read and write mechanism will cost 10% of what a tape drive costs.

Figure 4.8 compares the cost of reliability as we vary the time between successive generations of hard disks with removable platters. We also compare traditional hard disks as described in Section 4.4.1. We observe that separating hard disk platters from their read and write mechanism could cost significantly less than traditional hard disks in highly reliable archival systems, but the amount of the sav-

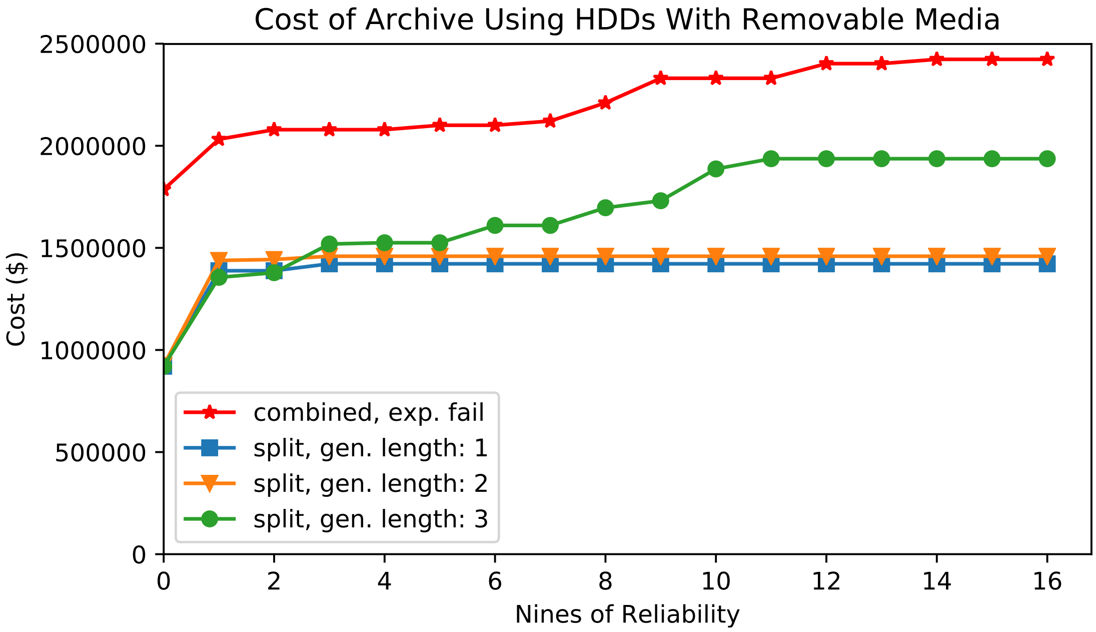

Figure 4.8: The cost of hard disks with separable platters in archival systems depends on how often the technology is updated with increased capacity and performance.

ings depends on how frequently the hard disk technology is updated. Updating the technology every one to three years could save between 42% and 20% compared with the cost for traditional hard disks. We conclude that exploring alternative design possibilities for established technologies like hard disk drives could result in meaningful savings for demanding and reliable archival systems.

4.4.5 SSDs for Reliable Archival Storage

Results in Section 4.4.1 showed that SSDs cost more than most other technologies for archival data storage. We explore the effects of increasing SSD capacity by considering the possibility that the current pace of SSD developments will continue further into the future and by examining the effects of increased AFR on the cost of reliable archival storage using SSDs. Figure 4.9 compares the cost for reliabil-

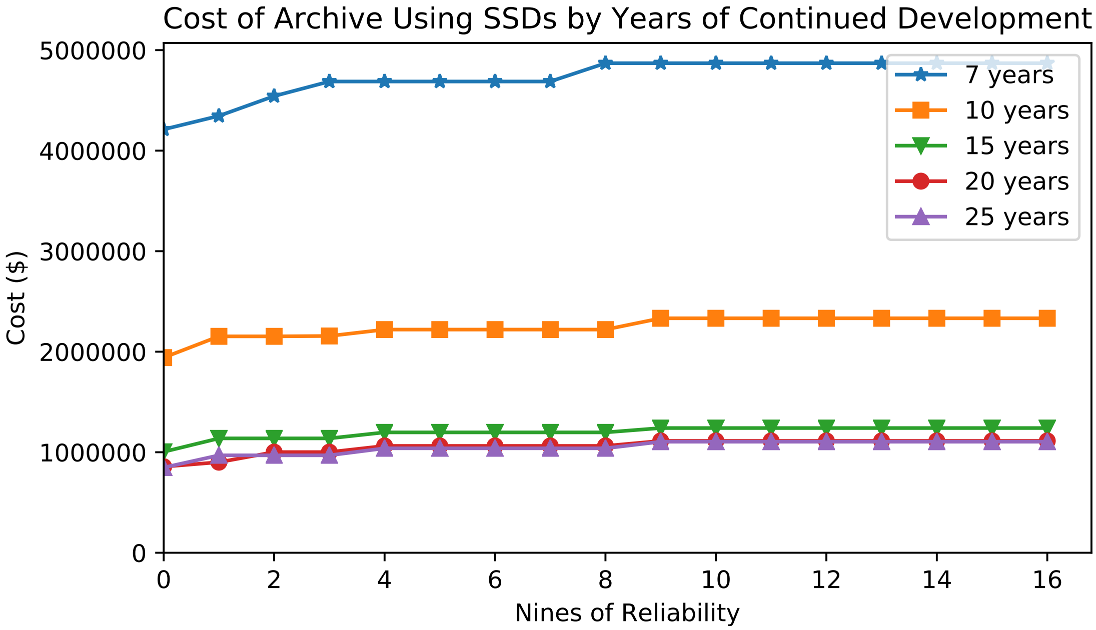

Figure 4.9: SSD capacity and development dramatically affect the cost of reliability in archival systems using SSDs.

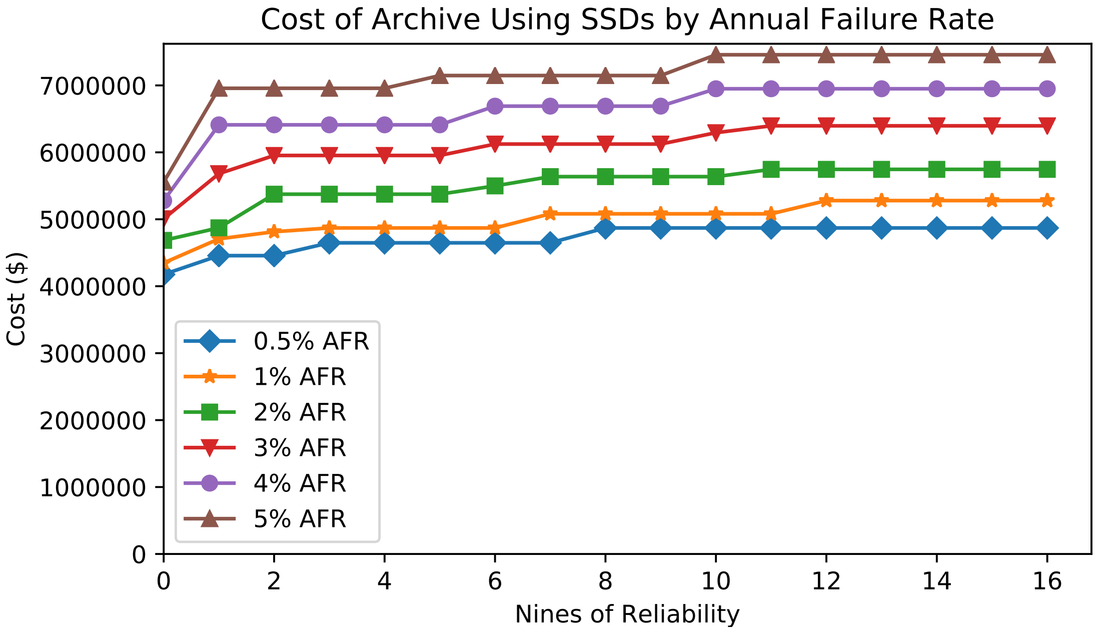

Figure 4.10: Higher AFR values have a marginal impact on the cost of reliability in SSD-based archival storage.

ity in archival storage if the development of SSDs continues apace for seven to 25 years. We observe that, as expected, a longer development roadmap, which would result in a lower cost per byte of SSD storage, reduces costs for SSD-based archival storage relative to a shorter development roadmap. We also observe that relatively short extensions of the SSD development roadmap can dramatically decrease the cost of SSD-based archival storage. Extending the development of SSDs apace for 10 to 15 years can reduce the cost of archival storage with 16 nines of reliability by 52% and 75%, respectively.

The emergence of QLC flash along with continued scaling and stacking of flash layers have increased capacity and reduced the cost of data storage in SSDs, yet such changes may come at the expense of SSD reliability. Some have argued that the lower endurance of novel SSD technology such as QLC outweighs its cost advantage and renders it unsuitable for archival storage [121], but the emergence of denser SSD technology may prove to offer cost advantages over less dense and, by extension, more reliable SSD technologies if the increased density does not prevent archival systems from offering a similar level of reliability at a lower total cost.

What effect would lower SSD device reliability have on the cost of archival system reliability over the long term? Figure 4.10 shows that large increases in AFR have a relatively small impact on the overall cost of reliability for archival storage with SSDs. Doubling the AFR from 0.5% to 1% increases the cost of archival storage by 8% over 25 years. Even if future developments to SSDs come at the expense of some device reliability, we predict that the increased capacity of such SSDs will notwithstanding make them ever more suitable for archival storage.

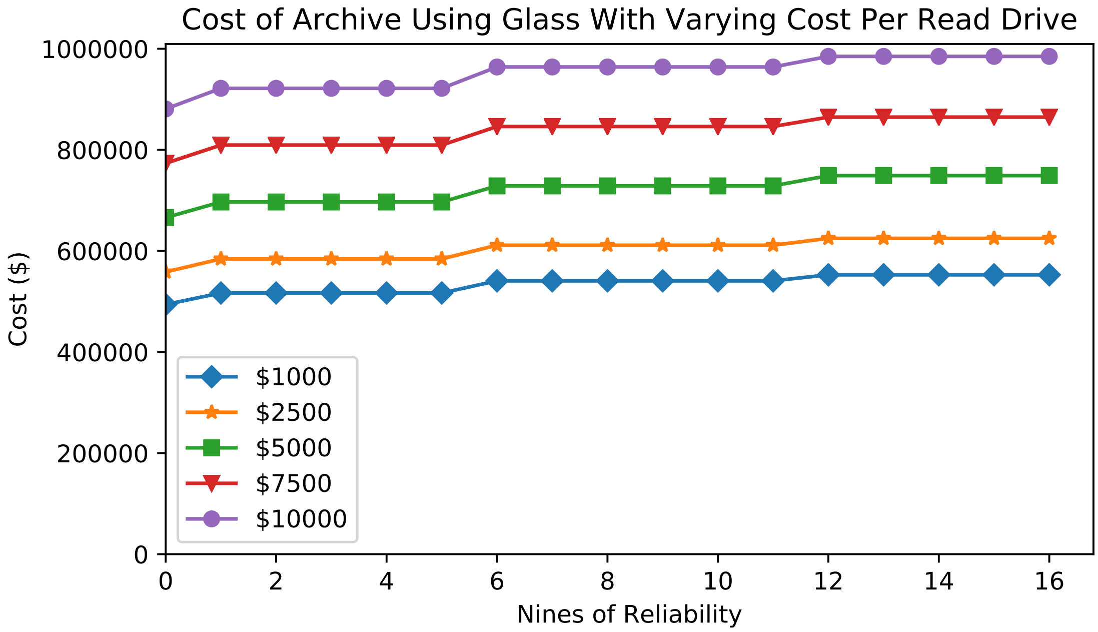

Figure 4.11: The cost of storing data in glass increases marginally with the cost of a reader drive.

4.4.6 Archival Glass

Archival glass promises to be a highly reliable storage medium, yet the exact cost of the hardware needed to read and write data into glass remains unknown. We designed experiments to measure the effect of increasing the cost of a drive to read glass from our baseline of $1,000 to $10,000, which is also the cost of the drive to write data in our experiments. We set the cost of media to $1 and its AFR at 0.01%. We choose a $1 cost for media to reflect the simplicity of the glass medium [8].

As shown in Figure 4.11, increasing the cost tenfold of a drive for reading data from glass increases the total cost of archiving data in glass by 78% for 16 nines of reliability. The total cost of reliability in glass storage thus increases only modestly with the price of its drive for reading because the intrinsic reliability of the glass medium requires only minimal parity to achieve high levels of reliability. We con-

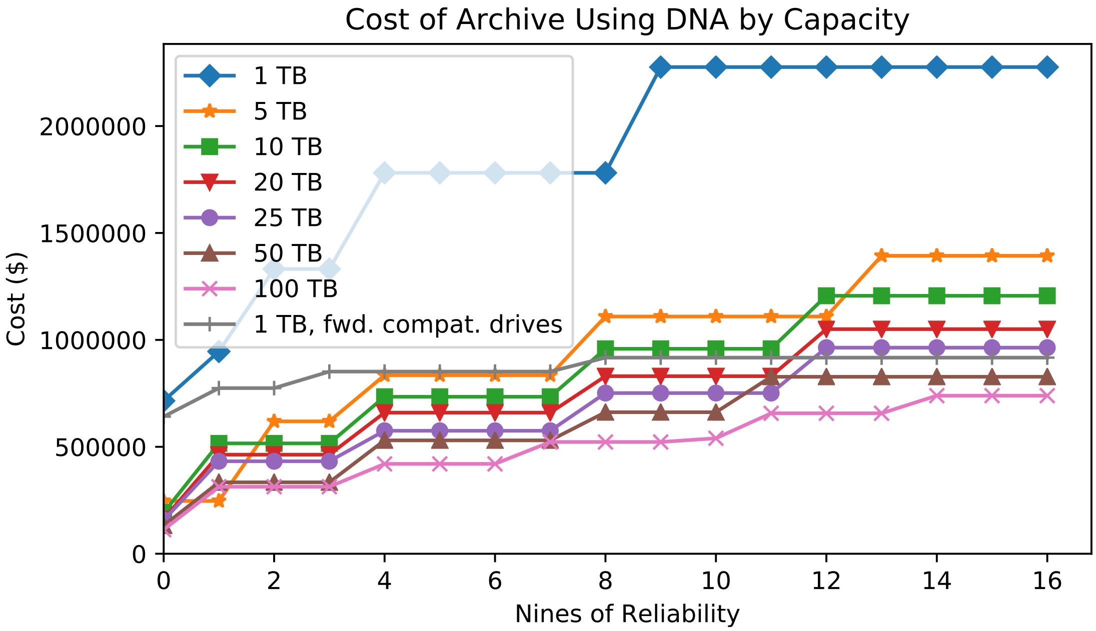

Figure 4.12: The cost of reliability in DNA-based archival storage depends upon the capacity and cost of each DNA molecule as well as the forward compatibility of DNA sequencers and synthesizers.

clude that glass has an advantage over other technologies due to its low cost and high reliability as a storage medium.

4.4.7 Synthetic DNA for Reliable Archival Storage

Synthetic DNA is currently in development, and we do not yet know how its development will proceed or if it is likely to provide the features that we have modeled. Although it is not our goal to predict the cost of individual DNA components and storage devices, we present these results to provide a baseline of performance and cost against which DNA-based archival storage systems can be compared and developed as time passes. We also leave it to continuing work to assess the real performance, cost, and reliability characteristics of DNA-based archival storage systems should they become commercially available.

Our previous experiment in Section 4.4.1 calculated the cost of DNA storage using a baseline of 1 TB per DNA molecule. We explore the effect on cost for reliability of increasing the capacity per molecule and, alternatively, envisioning DNA sequencers and synthesizers that can read and write any molecule of DNA produced in the future. We assume that each DNA molecule costs $100 in materials with an AFR of 1%, the sequencer costs $1,000, and the synthesizer costs $9,000.

Figure 4.12 shows that cost for each level of reliability decreases as the capacity of DNA increases. Cost for 16 nines of reliability decreases by 39% as capacity increases from 1 to 5 TB and by 68% with a capacity of 100 TB; however, enabling sequencers and synthesizers to read and write DNA molecules created with future generations of DNA technology reduces the total cost by 60% compared with our baseline that does not support forward compatibility. We conclude that flexibility in the design of DNA storage systems can help to dramatically reduce their cost for reliable archival storage.

4.4.8 Cost of Preserving Fixed Amount of Data

The demand for new advancements in hardware reflects the presumption that the demand for data storage is growing. If an archival system stores a fixed amount of data over a long period of time, then the controlling factor in the cost of the archival system becomes the initial acquisition cost of the system and the reliability of the storage devices within it. Figure 4.13 shows the total cost of an archival system implemented with each storage technology to preserve the same data over 25 years without adding to or modifying the data.

In this experiment, optical disc proves to be competitive with both tape and hard disk, particularly if we assume that hard disks exhibit an increasing probability of failure as they age. The stability and reliability of optical media also offsets its

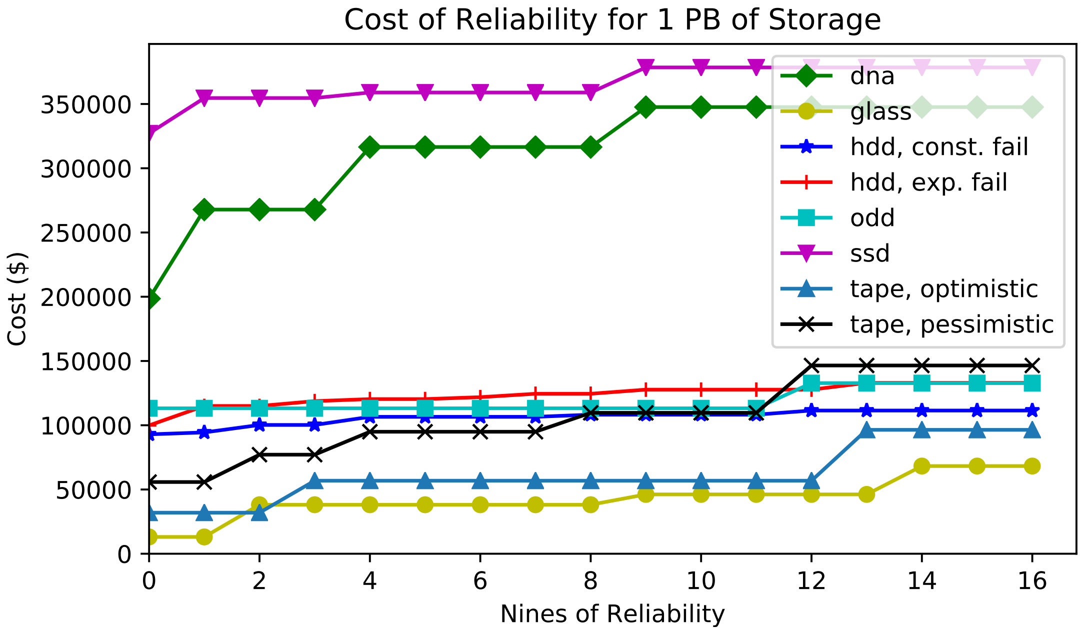

Figure 4.13: The cost of reliably storing 1 PB of data favors devices that offer high reliability.

limited prospects for development. Synthetic DNA and glass differ dramatically in terms of cost because of the disparity in our assumptions about the costs of their storage media, and our assumptions about the higher AFR of DNA compared with glass cause DNA to increase in cost more than glass as the storage system provides more nines of reliability.

4.5 Summary

Reliability is an important consideration for selecting and designing an archival storage system. Existing storage technologies are each capable of achieving many levels of reliability by incorporating the necessary amount of parity and redundancy, given the reliability of each storage technology; however, each existing storage technology has limitations that reflect the compromises inherited from its design. Novel storage technologies like glass and synthetic DNA may, if they become available in the future, outclass existing storage technologies over a range of possible use scenarios. In particular, the low cost and stability of archival glass media will allow glass to compete with tape for the lowest-cost technology for archival storage systems in the long term. Synthetic DNA may prove cost effective for highly reliable archival storage systems if, for instance, its design can relax the constraints of compatibility between different generations of technology. With such possibilities in mind, we expect that novel storage devices will outperform existing technologies for highly reliable archival systems in the future.

Chapter 5

The Cost of Workloads

Archival storage systems, although not often considered for their performance as much as for their cost, capacity, and reliability, meet the vital need for long-term storage that may nonetheless have significant performance requirements. We present empirical observations that confirm the intuition that archival workloads are both varied and often exigent. We present simulation data with a range of different workloads to explore the effect of workload on the cost of acquiring and operating an archival storage system. We evaluate our results to describe preferred workloads characteristics for each storage technology.

5.1 Introduction

Archival storage systems have long served the role of preserving information at a minimal cost over an extended time. As the amount of data in digital archives continues to grow, so must the capacity and performance of the devices and systems that store it.

Storage technologies that contend for applications in archival systems must keep pace with the ever-growing demands for capacity, performance, and reliability. Each storage technology offers unique features, limitations, and path to future development, and their unique capabilities may make them more or less suitable for a given archival system.

The suitability of each technology for archival storage depends upon how the technology's features and development roadmap align with the particular demands of an archival storage system. Existing storage technologies may suit some workload scenarios more than others since no single storage technology dominates all others in terms of cost per byte of storage, performance, and development roadmap. The design of archival systems using different storage technologies can greatly affect the cost, performance, reliability, and scalability of archival systems, and the variability of archival demands and storage technology developments further complicates the implementation of archival systems that both meet their users' needs and minimize long-term costs.

The relationship between archival workload and total cost of ownership has been defined only in the abstract. Intense workloads, it is reasoned, lead to greater total costs of ownership. Yet the reality of the relationship between workload and cost must be more complex than such cursory assumptions. Figure 5.1 illustrates a hypothetical relationship between two imaginary storage technologies, *Device A* and *Device B*. The figure maps the minimum cost to maintain an archival system over time with various workload conditions. The workload conditions, expressed here as annual read operations, affects how many storage devices must be used to serve the required workload. The storage technology with the lowest cost at a given simulation time and workload covers the corresponding location on the map. The map as shown illustrates an example of the transition between two technologies, and notably because different technologies develop at different rates over time, the line between the two device types may be non-linear. We can further complicate

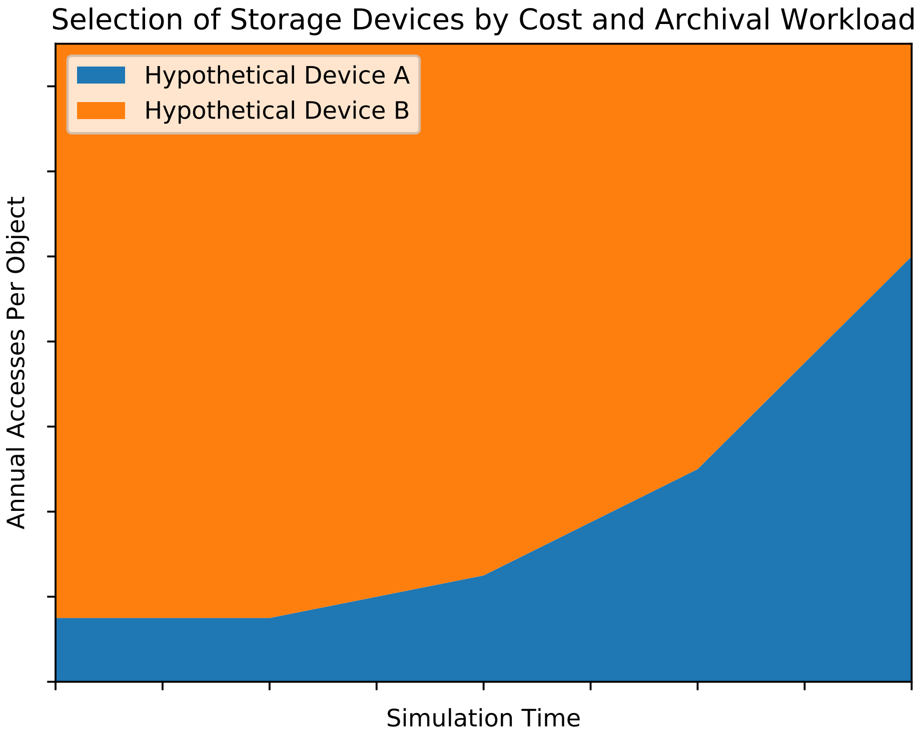

Figure 5.1: In this hypothetical example, we present a comparison of two device types *Device A* and *Device B*. The *y*-axis represents the intensity of the archival workload in the total number of read operations per year from the archival workload. The *x*-axis represents the time in the simulation, where time 0 is the beginning of the simulation, and the end of the axis is the end of the simulation. Each area on the plot shows which device, A or B, has the lowest total cost of ownership for any given workload intensity and time during the simulation.

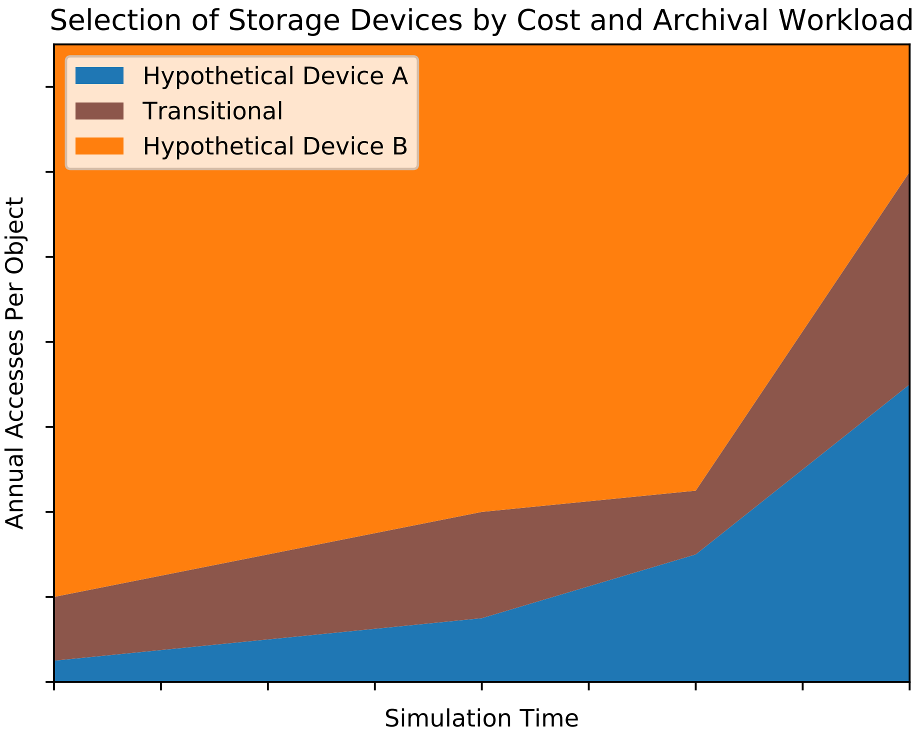

Figure 5.2: We show here a hypothetical example of the transitional space between the two device types as their total costs may prove very similar for such workloads during the simulation.

our picture of the relationship between cost and workload.

Figure 5.2 shows a similar map to that described above; however, here we show a transitional zone between the two devices. The transitional zone blurs the boundary between the two technologies in order to recognize the fact that our simulations return data that may vary between separate runs with identical information, and furthermore, the difference between two types of storage may be negligible within the transitional zone, falling within the bounds of a 95% confidence interval, for example.

The need to predict the viability of different storage technologies for archival storage systems emerges from the novel opportunity to define the boundaries of the relationship between candidate storage technologies for archival storage system and the workload that it must serve. In the following sections, we present details of our measurements and findings.

5.2 Simulator Setup

Our simulation model calculates the cost, performance, and reliability of an archival system by first taking as input a set of parameters for each simulation experiment. The parameters include the type of storage that is available for the simulation, the storage device cost, performance, reliability, power consumption, and other related features. The simulation parameters also include values for the amount of data that the archival system must store and the rate at which the data storage must grow to meet increasing demands.

We calculate the performance of the archival system in terms of both throughput and latency. Latency does not directly affect throughput as devices spanning several orders of magnitude from microseconds in the case of SSDs to minutes in the case of Tape may deliver similar levels of throughput; however, latency can reduce the effective throughput of any storage technology by increasing the amount of time required by each storage device before it can begin to operate at its design throughput. Latency then can affect the total throughput that a storage system may achieve by limiting the amount of data that it can access over any given length of time.

We calculate total throughput as the maximum combined throughput of storage devices available for reading or writing data within an archival system. The total throughput available to read a single piece of data depends upon the performance of the individual device or devices where that data is stored. Throughput in the context of our simulations is thus a measure for performance that scales with the number of storage devices in the storage system, and the total system throughput can grow by adding more drives to read or write data. At the same time, the growth of total throughput does not necessarily increase the throughput available to read any singular data object, as that throughput depends on individual devices. Storage technologies with removable storage media such as tape, glass, and synthetic DNA can still scale the combined throughput of multiple tape drives so long as the necessary infrastructure exists for writing data to multiple devices simultaneously.

In addition to total throughput, we also calculate the latency for reading and writing data into the archival system. Unlike throughput, the performance measure of latency does not scale as more drives are added to the storage system. A tape-based archival system, for instance, requires a fixed amount of time to load a tape cartridge into a drive, and the latency of accessing a tape does not decrease as the total system throughput increases; however, alternative storage technologies have the potential to offer better performance than others in terms of throughput or latency. Taking the technological features of each storage device and the simulation parameters together, the simulator searches for a configuration of storage devices that meets the requirements for capacity, reliability, and performance at the each time during the simulation.

5.3 Experimental Setup

We set up the simulator to measure the effects of different workloads upon archival systems using any one of the storage technologies that we consider. We measure workloads in terms of the total number of read operations per year since, as noted in Chapter 2.1, archival storage systems follow a write-once-read-maybe pattern. The total number of write operations for archival data must meet the demands to write the data into the archival system, but once written, the variable number of reads can prove to be either small and possibly insignificant or large and possibly decisive.

We configure the workload parameters of the simulator to run simulations with read workloads between 0 and 10 million read operations annually, or approximately equal to between 0 and 0.3 operations per second in the archival system. Since some storage technologies like SSD can respond quickly to random-access workload requests while other devices like tape respond more slowly with higher latency, the simulator must scale the archival system appropriately by adding a sufficient number of drives to meet the demand of a the archival workload. For this reason, we modified the simulator to add more drives when needed to meet the demands for additional throughput and read operations.

Each device type characterized in our simulation model includes as part of its parameters, values that describe its throughput on read and write operations and also its latency. We compare these values with the workload parameters for each simulation, and we vary the workload parameters between 0 and 10 million read operations annually to explore a wide range of workload demands. The simulator calculates how many drives of a particular type will be required to achieve the performance threshold from the workload parameters. We can calculate the number of drives needed for a particular workload as

$$T_d = \frac{\text{Size}}{R_d} + L_d + A_d + U_d, \tag{5.1}$$

where T_d is the time on average used by the drive to read each data object, Size is the average size of data objects in the workload, R_d is the read throughput of the storage device, L_d is the time to load or spin up a drive from idle, A_d is the average latency off the average latency of the device when seeking for data, and U_d is the time to unload a drive if necessary. We use a similar formula with appropriate parameters for write operations. We can then calculate the number of drives needed for a particular workload as

$$N_d = \left\lceil \frac{T_d \times \text{Ops}}{\text{Len}} \right\rceil, \tag{5.2}$$

where N_d is the number of drives needed, Ops is the number of read operations in the workload measured annually, and Len is the length of a year measured in seconds. We measure the length of a year in seconds because the throughput of each drive is measured also in seconds. We also take the ceiling in Formula 5.2 because the performance of the storage system must meet or exceed the workload specification. After calculating N_d, we proceed to add drives to the simulated archival system to meet the workload demand. Although Formula 5.2 affords us the ability to estimate the number of drives required to meet a given workload, there remain relevant albeit peripheral factors that may affect the performance of the storage system.

In calculating the number of drives needed to serve a given workload, we have included the principal factors that affect the total performance of a storage system; nevertheless, there are other factors that can impact the performance of the storage system. Foremost among these is that the passage of time often leaves older devices with somewhat slower performance over their lifetimes while new devices have higher performance. The total throughput of the storage system will also be affected by external events like power outages, device failures, and maintenance. Our simulator as described in Chapter 3 tracks delays caused by device failures and the performance of new drives; however, we limit the scope of our work to those factors that derive their importance from the design of the storage devices themselves, including throughput, latency, and reliability, and we rely on the simulator to expose the relationships between each factor through the results that it produces.

5.4 Experimental Results

The definition of archival storage allows for the possibility of a wide range of workloads that vary in intensity. We measure the intensity of an archival workload in terms of its annual read operations, and we configure our experiments relating to the cost of archival workloads by varying the number of read operations that the archival systems serve. Our experiments vary the annual read operations between zero and 10 million to explore the effects of various workload intensities upon different storage technologies.

5.4.1 Workload and Total Cost

Figure 5.3 shows the results of simulations for the total cost of archival systems constructed with different storage technologies, including both existing and prospective. As before, we consider the prospective technology of archival glass in lieu of the optical disc to evaluate the potential for optical storage technologies. Synthetic DNA as a storage technology proves to be competitive with other storage technologies for workloads with the lowest intensity, but the cost of DNA rises linearly with the demands of the workload it serves. The growth of cost for DNA-based archival storage suggests that synthetic DNA can best serve those archival systems that require large capacity on the one hand and small workload of read operations. Nevertheless, as time passes, there may prove to be more opportunities for synthetic DNA relative to other storage technologies. The cost of other storage technologies as shown in Figure 5.3 cluster in the lowest part of our simulation ex-

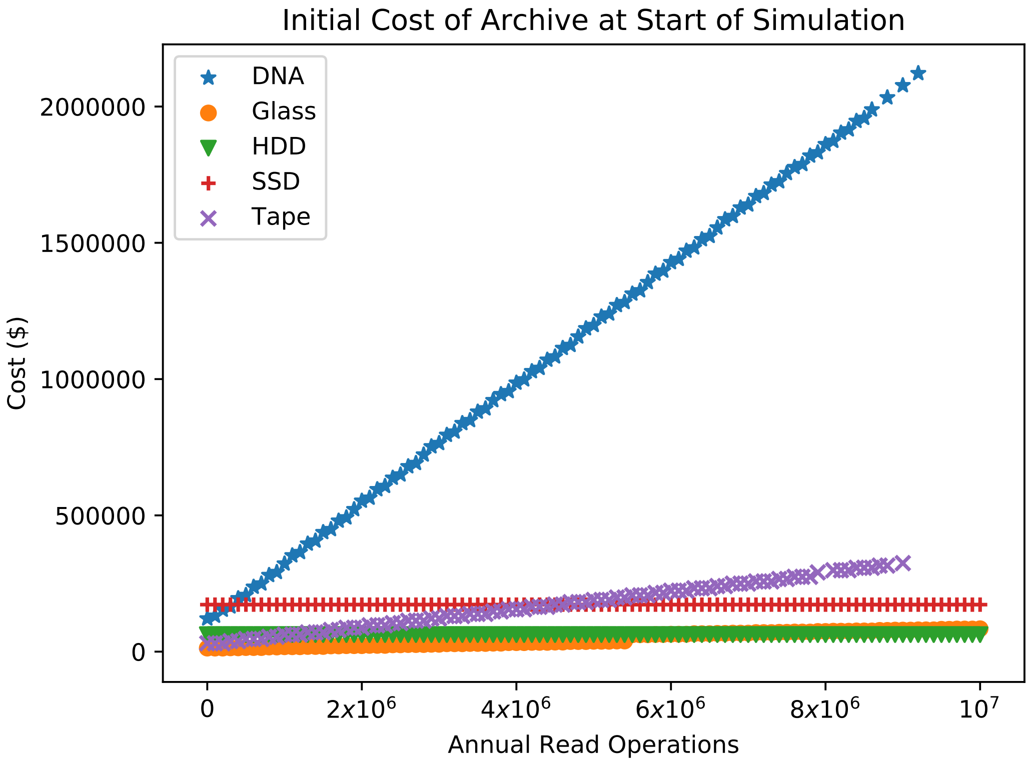

Figure 5.3: The cost of archival storage depends upon both the storage technology in use and the workload of the archival system. As different technologies feature their own levels of scalability, the effects of increased workload differ between storage technologies. Here, we show the cost of archival storage from the beginning of the simulation. We plot the annual workload of the archival system in terms of annual read operations on the *x*-axis and total cost of ownership on the *y*-axis.

periments, and we plot the same data again with only the first $500,000 of cost on the *y*-axis in Figure 5.4.

Figure 5.3 and Figure 5.4 show the state of storage technologies for archival storage relative to workload roughly as they exist at the current time, but in the case of prospective storage technologies, we rely upon expectations for cost and performance derived from published literature. Figure 5.4 in particular presents the relative cost of each storage technology for a wide range of archival workloads. We expect that, although traditional archival storage technologies like tape and hard disk provide sufficient capacity to meet current demands, the effects of increasing workloads causes certain technologies to become uneconomical relative to others. Hard disk drives and solid state disks, for instance, do not cost significantly more to operate with even the most intense archival workloads that we evaluate. Tape, on the other hand, becomes less economical as the workload's intensity increases. Tape is the most economical of existing and traditional storage technologies for workloads up to 1 million annual read operations, but hard disk drives prove to be more economical than tape for all workloads over 1 million annual operations. Additionally, it becomes more economical to use solid state disks instead of tape for archival storage systems that demand over four million annual read operations. The performance of SSDs can therefore accommodate both a high-performance and intense workloads as well as a moderate workloads while still costing less than tape for the near-term at least. Solid state disks today have greater flexibility than does tape for workloads that straddle the definition of archival and primary workloads.

Considering our definition of archival storage from Chapter 2.1, archival storage systems imply association with low-intensity workloads, but there is no clear boundary between archival and non-archival workloads. With such a definition in mind, we may consider workloads to be more characteristic of archival as their read intensity approaches zero. Similarly, we may also consider a workload to be less characteristically archival as its intensity of annual read operations increases. The lack of clear boundaries between archival and non-archival workloads reflects the possibility that some archival storage systems may be used nearly as intensely as primary storage systems [3, 129, 141]. Such archival systems would, at least hypothetically, benefit from the use of storage technologies that can easily support high-intensity workloads with a minimum of cost inflation over their use with only

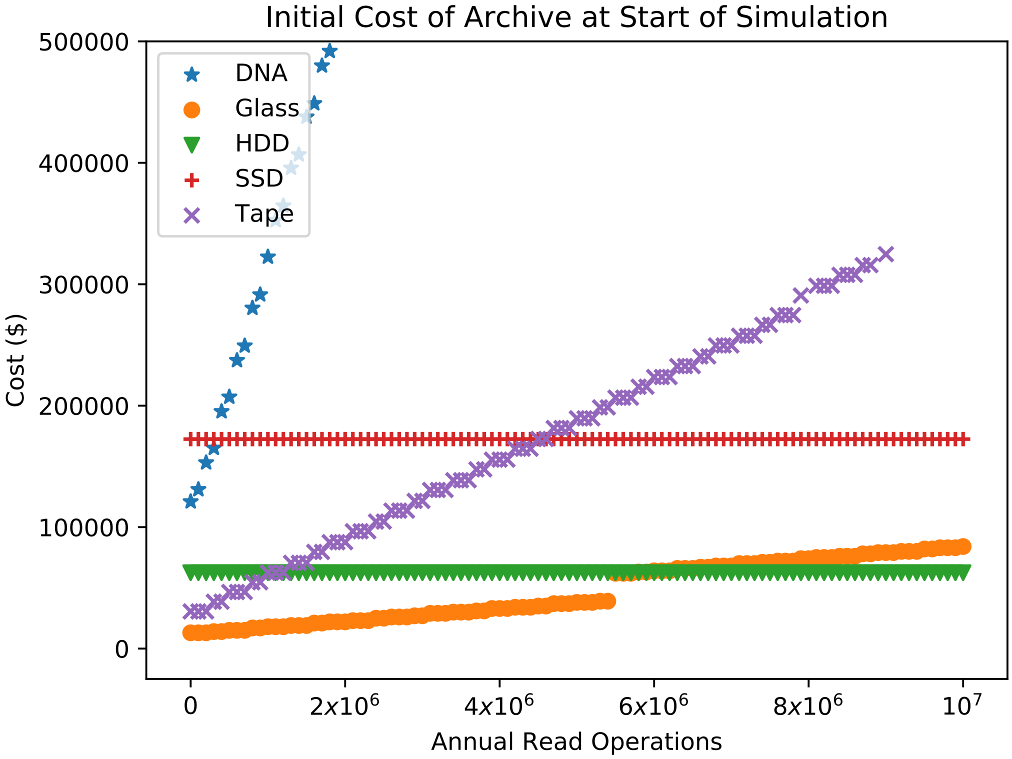

Figure 5.4: The first $500,000 of simulated cost on the y-axis reveals in greater detail the relationships between existing and prospective technologies for a variety of archival workloads. The cost of glass is lower than other technologies for workloads with fewer annual read operations, but hard disk becomes more economical as the workload's intensity increases.

a small number of annual read operations. Figure 5.4 shows that tape becomes less economical as the workload becomes less characteristically archival with growing intensity. Archival glass, which glass, which also employs separable media and drives, suffers comparatively less than tape does with more intense workloads since the cost of glass-based archival systems will exceed that of hard disk drives above workloads of approximately 5 million annual read operations. We can extend our observations to the end of our simulation time of 25 years to explore how our expectations change with time.

Figure 5.5 shows the total cost of archival storage systems with different storage technologies at the end of the simulation time of 25 years, varying only the total number of annual read operations. As with our findings at the beginning of the simulation, synthetic DNA proves to be competitive with both tape and archival glass for the coldest workloads near zero annual read operations; however, its economical attractiveness decreases linearly as the workload grows more intense. Synthetic DNA passes the cost of hard disk drives near four million annual read operations. Tape, although less economical than glass for cold workloads, grows even less competitive relative to glass as the workload demand increases. Tape and glass both feature removable media, so we should expect that they would each suffer a growing disadvantage with intense workloads relative to technologies like that have combined media with drives, namely hard disk and solid state disk. Yet even though their cost grows over time, the upper limit of our simulations with 10 million annual read operations is not nearly enough to cause either tape or glass to exceed the cost of hard disk or solid state disk.

We saw earlier that hard disk and solid state disk can be competitive storage technologies for archival systems, and that they can deliver grater flexibility since they can easily serve intense workloads for primary storage while still delivering lover total cost of ownership than tape, glass, or synthetic DNA for certain archival workloads. Figure 5.5 shows that the role for hard disk drives and solid state disks, however easily they accommodate workloads ranging from cold to hot, will shrink as the growth rates for their capacity decreases in the coming years and decades. The demand for archival storage capacity and throughput will certainly outpace the growth rate of HDD and SSD capacity, and the performance advantages of HDD and SSD cannot compensate for their anemic capacity growth rates, particularly when compared with the potential of prospective storage technologies like glass and, to a lesser extent, synthetic DNA. The value of HDD and SSD technol-

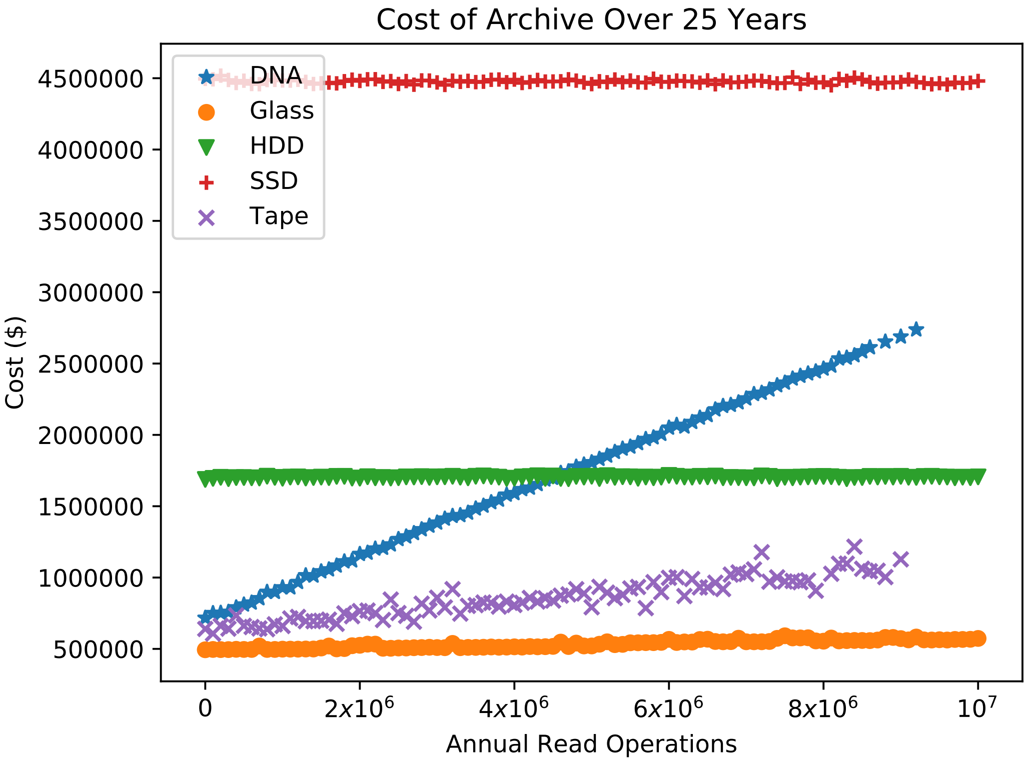

Figure 5.5: The cost of archival systems for a range of workloads changes after 25 years of simulation time. The competitiveness of SSD and HDD decrease relative to other technologies, while the competitive advantages of prospective technologies glass and synthetic DNA continue to strengthen. Tape continues to be a viable option for archival storage systems well into the future, but its scaling limitations leaves an opportunity for other technologies to serve high-intensity workloads more economically.

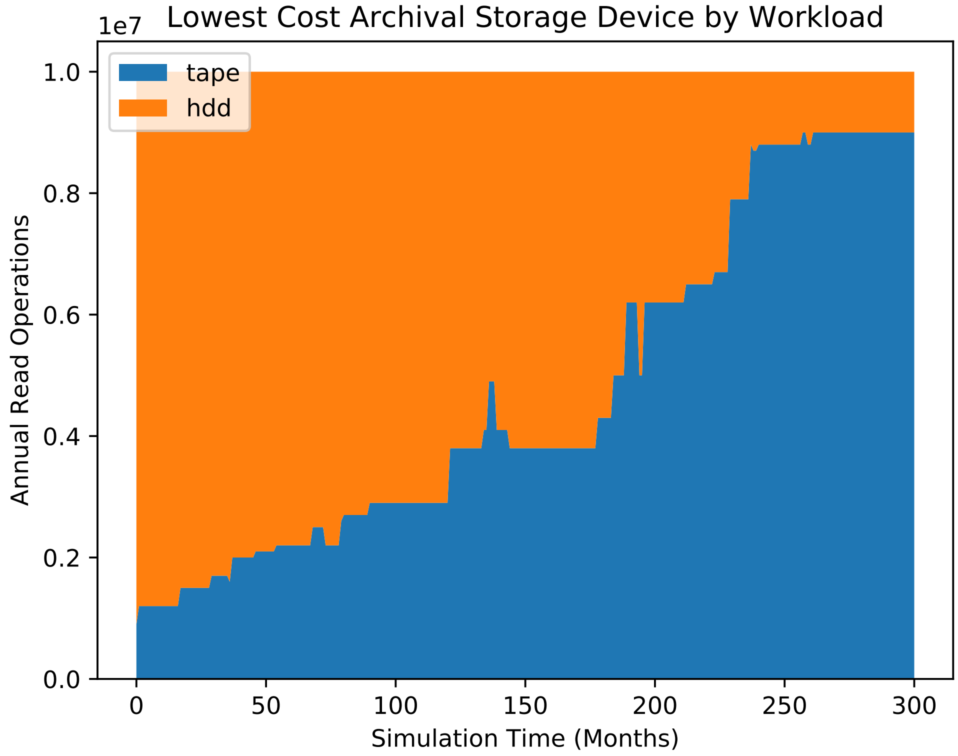

Figure 5.6: We compare hard disk and tape over the 25 years of simulated time by searching for the lowest-cost device type over a range of workload intensities and as time passes. Here, we show workload in the *y*-axis and time in the *x*-axis.

ogy for archival systems may nevertheless remain relevant if, for instance, particular archival systems require lower latency than either DNA, glass, or tape can deliver. We expect therefore that the future of HDD and SSD technologies will increasingly relegate them to the niche of high-performance and low-latency archival workloads.

Hard Disk Drives

Figure 5.6 presents a comparison of hard disk and tape based on their total cost of ownership. We color areas on the figure for either tape or hard disk depending on which device type costs the least for the combination of workload intensity and simulation time. We observe that tape, as expected, dominates the top of the figure that corresponds to the more demanding workloads of annual read operations, and tape dominates the bottom of the figure where the workloads are the coldest. As time passes, we see that tape takes over more workloads where previously HDD had dominated. The declining role of hard disk relative to tape over time reflects the narrowing niche available for HDD in the future due to its likely future of slowing development.

Although the long-term future of HDD technology for archival storage will become more tenuous over time, it still remains true for now and at least a decade into the future that hard disk drives will be more economical than tape for many workloads. After the next decade, new technologies like archival glass and synthetic DNA may become commercially available for widespread adoption, and the limited lifetime of HDD-based archives could facilitate a fortuitously anticipated migration to a new, cheaper, and more scalable storage technology like glass.

5.4.2 Workloads and Cost Scaling

Archival systems must be designed for maximum flexibility as they meet the demands of their designers for capacity, reliability, and performance. Flexibility is important because it ensures the storage system can perform adequately even when serving unanticipated needs, including more demanding workloads than that for which it was designed. Each storage technology that we have considered offers some baseline of performance, and some technologies can support much more throughput relative to their capacity than others. Archival storage systems that utilize hard disk or solid state disk, for instance, typically scale their performance at the same time as they grow their capacity since each drive delivers its own capacity and throughput. Tape, glass, and synthetic DNA, on the other hand, add storage capacity separately from throughput, and glass and DNA even add read and write throughput separately. For these reasons, some technologies will serve workloads with many read operations with almost the same cost as they will serve workloads that have zero read operations.

Figure 5.7 and Figure 5.8 show the scaling of cost for each storage technology, normalized to the total cost of ownership for the workload of zero annual read operations. Synthetic DNA exhibit by far the greatest degree of cost scaling as we require of it more annual read operations. Tape and glass also exhibit cost scaling; however, the scaling of glass remains lower than that of tape both at the beginning and end of the simulation time. Glass enjoys a lower scaling factor for more demanding workloads than does tape due in large part to the design of the glass archival storage technology. While tape drives perform both read and write operations with the same device, glass storage splits its workload between writer drives and reader drives. Glass drives that read data can do so while writer drives save new data into the archive, a feature that introduces intrinsic parallelism into the design of archival glass storage systems. Additionally and most crucially, the cost of glass reader drives is far lower than the cost of either glass writer drives or tape drives. Glass can therefore add more reading throughput for less marginal cost than can tape-based archival systems. Glass shows a large discontinuity in Figure 5.7 where the scaling of cost jumps dramatically around 5.5 million annual read operations. The discontinuity is the result of the need to add an additional library for the glass storage system, and the library has its own costs in addition to the need for matching writer and reader devices. The unique event that causes the discontinuity at the beginning of the glass simulation disappears in Figure 5.8 for the end of the simulation time because simulations for each workload level added new libraries to accommodate the demand for additional capacity, and thus the differences be-

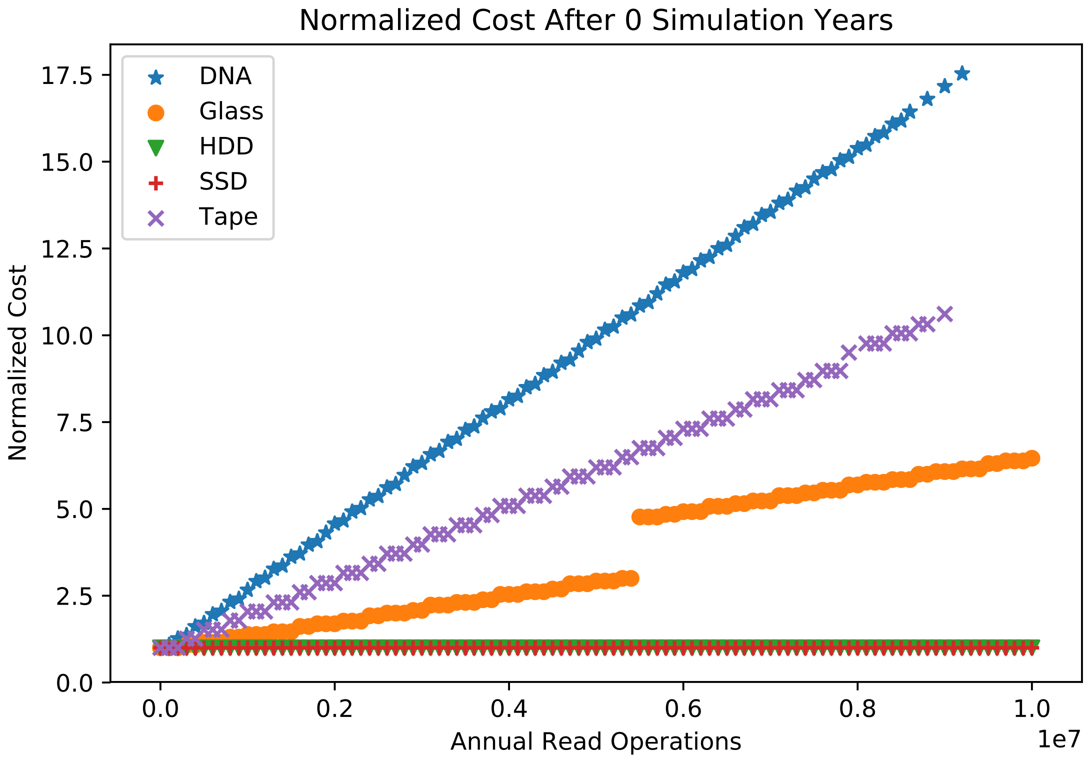

Figure 5.7: We normalize the cost of archival workload data by dividing each storage device with its own cost to serve a workload of zero reads. Device types that can serve larger workloads at little or no additional cost relative to their base case present flat lines without. Device types for which demanding workloads require additional costs present positive-sloped lines as their cost grows with workload relative to the base case of zero reads.

tween the separate simulations for glass was how many reader drives each simulated archival system needs to meet its workload read demand. Unlike glass, hard disk and solid state disk exhibit their own cost scaling pattern that typifies their design characteristics.

HDD and SSD devices feature designs with combined storage media and reader mechanisms. The integration of the reading mechanism into the storage device results in a close relationship between the number of storage devices in any system and the total throughput available from the drives in that system. Archival storage systems with HDD and SSD devices can deliver throughput that scales with capacity, assuming that the drives can be used in parallel with one another and are not limited by heat or power constraints [12, 18]. Figure 5.7 and Figure 5.8 both show that HDD and SSD exhibit almost no scaling of cost as the workload increases in the archival storage system. Indeed, the only cause of scaling in such systems is from the additional electricity used by the more active storage systems. Although the cost of using glass and tape for archival storage can prove lower than the cost of SSD or HDD for archival storage, the flat scaling of these technologies shows that archival systems using HDD or SSD can easily adapt to the possibility of growing workload demands without suffering either inadequate performance or the need to purchase additional hardware. HDD and SSD fill the niche of flexibility which, however narrow it may prove to be, delivers important value for archival storage systems.

5.5 Conclusions

We have discussed the role of workload in archival storage systems, and we have presented data that shows how different storage technologies respond to various workloads. Existing storage technologies will struggle to adequately serve the demand for archival storage, particularly if that demand accompanies unplanned workloads with elevated levels of data accesses. The advantages of each existing storage technology diminish as their development roadmaps fall behind the demand for archival capacity and throughput, but prospective storage technologies will deliver solutions for the growing needs of archival storage systems, even for archival storage systems with demanding workloads. We found that Synthetic DNA can have a viable role within archival systems when utilized with the coldest archival data that is rarely if ever accessed. Glass in particular promises to deliver good

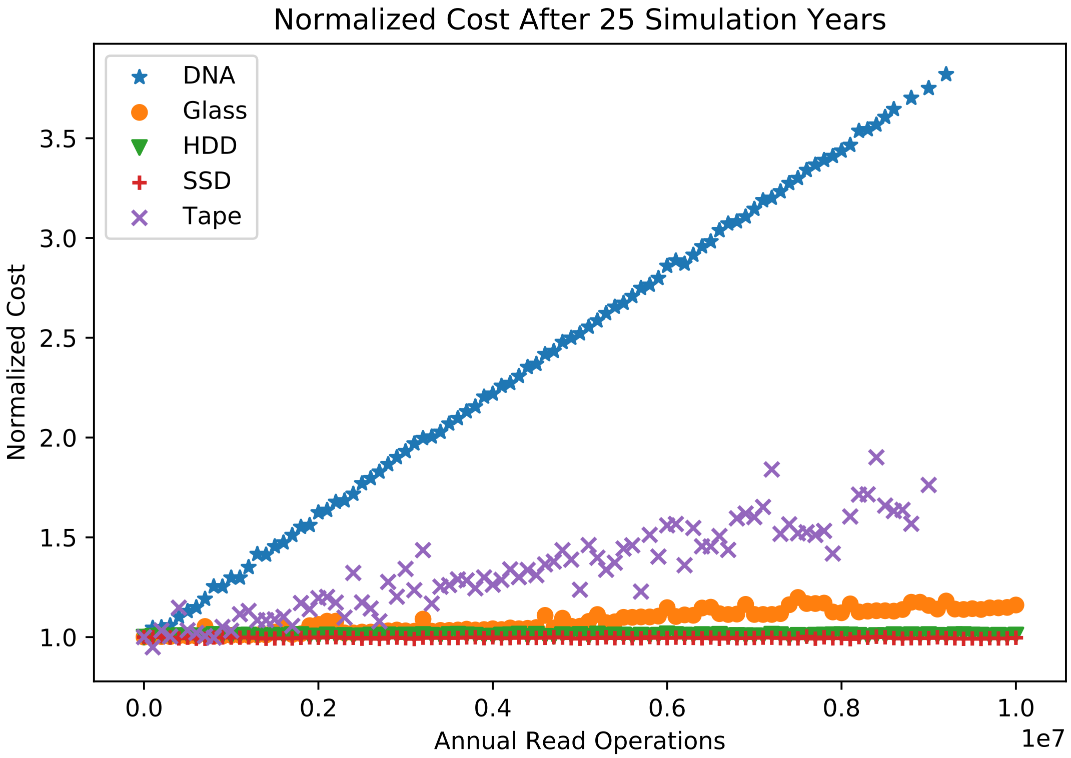

Figure 5.8: The normalized workload at the end of our simulation time, 25 years, shows the effects of read workloads upon different storage technologies in the long-term.

performance, modest cost scaling for demanding workloads, and a long future of potential improvements to maintain its cost advantage relative to tape and other storage technologies. Glass in particular will prove emerge as default archival storage technology that will relegate other technologies like hard disk and solid state disk to their respective niches within the archival storage market.

Chapter 6

Validation of Simulation Model

We have described the long-term economics of existing and prospective archival storage technologies with respect to their reliability, workload demands, and their prospects for further development. We have also explored the ways in which alternative scenarios can affect their competitiveness in archival storage systems. We also described the inner working of the simulation model that we use to generate our results. In this chapter, we seek to validate our simulation model by comparing its predictions with those of other models for evaluating the long-term economics of archival storage.

6.1 Introduction

Our simulation model has allowed us to explore the effects of different hypothetical trends upon the the economics of archival storage. Until now, we have compared the results of different experiments with each other to gather an understanging of how changes to storage technologies or the way in which they may be used will affect their relative cost over time. Earlier chapters thus facilitate a relative comparison between different storage technologies. Validation in the context of our simulation model will allow us to compare the results that we generate with those of other models or findings.

There has been no previous work that has exactly matched what we have achieved with our archival simulation model. Although this work is unique in its scope, goals, and approach, we can indeed find other works where some predictions have been made and justified concerning archival storage, and by comparing the results of our simulation model with the results from other works, we can at least enumerate how precisely our results match with the results in other works. To do this, we will also need to match the scope of our validation to the experiments and scope of other works. We begin validating our simulation model first by introducing other models that measured the relative economics of archival storage with different approaches and goals.

6.2 Validation Data Set

Our approach to validation—comparing our results on a given set of inputs with the results from other research on identical inputs—relies upon the available data about the exact parameters and results—inputs and outputs—from other work. The data set that we use for comparison is from the earlier work of Jeff Inman et al. that compared the cost of tape with that of hard disk drives for an archival storage system at Los Alamos National Laboratory (*LANL*) [57]. The data set includes historical information about the requirements and parameters for an archival system between 2002 and 2008. It also includes parameter to feed a linear model for estimating the cost of an archival system between 2012 and 2025 [42]. We examine each data set in the following sections, comparing them with our own parameters and results.

6.2.1 Archive Parameters

We begin first by introducing the parameters from the validation data set. The validation data set, although narrowly targeting the specific conditions of Los Alamos National Laboratory, contain some of the features that should by now seem familiar for modeling archival storage. Table 6.1 records the values that the reference data set presents for its historical model for years 2002 through 2008. The requirements for both data capacity and throughput grow much more quickly than the assumptions that we have utilized in our simulations. In particular, the growth of archive data for the historical validation amounts to a 142% CAGR, which is far higher than the 30% CAGR that we have used in simulations as described in Chapter 3. Additionally, the growth rate for the archive's data grows more quickly during some years than others. The uneven rate of data growth can have a nonnegligible effect on the archive's total cost if, for example, the greatest growth of data must occur during a time when the capacity of storage devices has not increased. The LANL data set thus introduces the risk that the demand for storage capacity may increase at an inconvenient time when the chosen storage technology has not recently delivered an increase in capacity. The same risk may also apply for throughput and other performance metrics, should the demand for throughput grow inconsistently over time.

The growth of the throughput demand for the validation data as shown in Table 6.1 grows less quickly than does the demand for capacity, albeit still more quickly than the baseline of our experiments from earlier chapters as described in Chapter 6.1. The faster growth of demand for capacity than throughput reinforces the notion that archival storage systems are characterized by their demand for large capacity more than they are characterized with their demand for high throughput as described in Section 2.1. Although the validation data set presents different requirements for archival systems that do our earlier assumptions, the opportunity to evaluate the output of our simulator using a novel scenario will demonstrate both the precision and the adaptability of our simulation model for real-world archival systems. We therefore use the values in Table 6.1 within the archival parameters for the validation of our simulation model.

6.2.2 Storage Device Parameters

The validation data set from Los Alamos National Laboratory includes values for device performance that differ from those that we have used elsewhere in our simulation model. In particular, the device parameters in the LANL data set model those from Oracle's StorageTek tape media, drives, and library systems. We present the device parameters in Table 6.2. The validation data set utilizes a two-year cadence between generations of StorageTek devices, and each new generation doubles the capacity and throughput of the previous generation, a feat that the LTO Ultrium devices from our other experiments have not achieved [21, 50, 51]. We utilize device parameters from the LANL data set in our simulation model for our validation simulations.

6.2.3 Device Numbers and Cost

The LANL data set includes values for how many tape drives, tape cartridges, and tape library systems serve the growing demands of the model archival system. We show the number of devices for each year in Table 6.3. The tape drives and media as shown span four generations of technology, with each new generation roughly doubling the one that came before in terms of performance or capacity. Each generation lasts two years, with a doubling of capacity and throughput available from a new generation every other year. The LANL data set begins in the mid-

Table 6.1: Requirements for Archival System

Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**CAGR**
Capacity (TB)	500	1500	4000	8000	20000	50000	100000	142%
Throughput (GB/s)	1.35	3	4	6	10	15	20	57%

Table 6.2: Tape Device Parameters

Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**CAGR**
Generation	1	2	2	3	3	4	4	-
Media Capacity (GB)	30	100	100	200	200	400	400	54%
Media Cost ($)	50	70	70	100	100	120	120	16%
Drive Throughput (GB/s)	0.015	0.03	0.03	0.06	0.06	0.12	0.12	41%
Drive Cost ($)	25000	30000	30000	30000	30000	30000	30000	3%

dle of the first generation so that second generation tape devices are available on the second year of the model. The shortening of the first generation of storage technology and early availability of the second generation many help to reduce costs in the LANL model. The doubling of capacity and throughput every other year is somewhat faster than the baseline from other experiments relating to tape since, as described in Chapter 3, the LTO tape specification delivers a new generation every third year. We nevertheless utilize the LANL device numbers in simulations for the purposes of validation.

Table 6.3 presents a mostly predictable pattern of device acquisition for any growing archival storage system with tape; however, one change in particular suggests the possibility that unexpected changes may lead to almost inexplicable decisions. The growth of data and throughput demand leads to ever-greater numbers of drives and tape cartridges, always adding the latest generation of tape drives and media. Generation three of tape technology begins as the other generations do, but once the fourth generation of tape technology becomes available in 2007, Table 6.3 shows the removal of all third generation drives and media in favor of a relatively smaller number of generation four devices. Interestingly, however, Table 6.3 records that no devices in the first or second generation retired from the archive at any time within the available data. The high cost of libraries in the LANL data set supports the early retirement of older devices in favor of a smaller number of newer ones, but we would expect that the oldest generation of technology should retire before a newer generation. Nevertheless, there are other explanations for why the third generation of technology would retire from the archival system early. Among these reasons are the possibility that third generation devices proved to be unreliable in practice, that their manufacturer withdrew support prematurely, or that all those devices were damaged or rendered inoperable as the fourth generation of devices became available. Any of these reasons exists outside the scope of both the LANL model and our simulator, but we must acknowledge external and unforseen eventualities as potentially disruptive for archival storage systems or leading to unforeseen costs.

We use the data from the LANL model to create parameters for our own simulator, and we compare the output of our simulator with the LANL model to gather insights about how our own simulation model relates to other approaches.

6.3 Comparison and Discussion

We present data comparing the output of our simulation model with the LANL model in Figure 6.1. We show the LANL results next to three separate simulations. With the CAGR growth model, we utilize the average CAGR growth rate of 54% for tape cartridge capacity, taken from Table 6.2. With the Step Growth models, we use the exact values for device capacity from Table 6.2 since the growth of capacity between the first and second generations is larger than the growth between other generations of tape. Finally, we vary the number of annual read operations for the archival system. The number of annual read operations was not included in the model data set, but varying the workload affords us the ability to compare its results with that of the LANL data. We run each simulation ten times and plot the average cost of each run together with the 95% confidence interval. Figure 6.1 shows in particular that the cost of the archival system in the LANL data set grows more slowly than does the data within our simulation model. The simulation model we use includes factors such as the cost of electricity and device failure and replacement, but the LANL data set does not include such values. The effect of power cost and device failure remain small due to the short length of time in the simulation—6 years in total rather than 25 years in other simulations from earlier chapters. Most

Table 6.3: Total Number of Tape Media and Drives By Year

Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Media Gen. 1	16,666	16,666	16,666	16,666	16,666	16,666	16,666
Media Gen. 2	-	10,000	35,000	35,000	35,000	35,000	35,000
Media Gen. 3	-	-	-	20,000	80,000	0	0
Media Gen. 4	-	-	-	-	-	115,000	240,000
Drive Gen. 1	90	90	90	90	90	90	90
Drive Gen. 2	-	55	89	89	89	89	89
Drive Gen. 3	-	-	-	33	100	0	0
Drive Gen. 4	-	-	-	-	-	92	134
Libraries Required	3	5	9	12	22	28	49
Annual Cost ($)	**3,383,300**	**2,550,000**	**3,170,000**	**3,290,000**	**9,010,000**	**17,160,000**	**18,360,000**

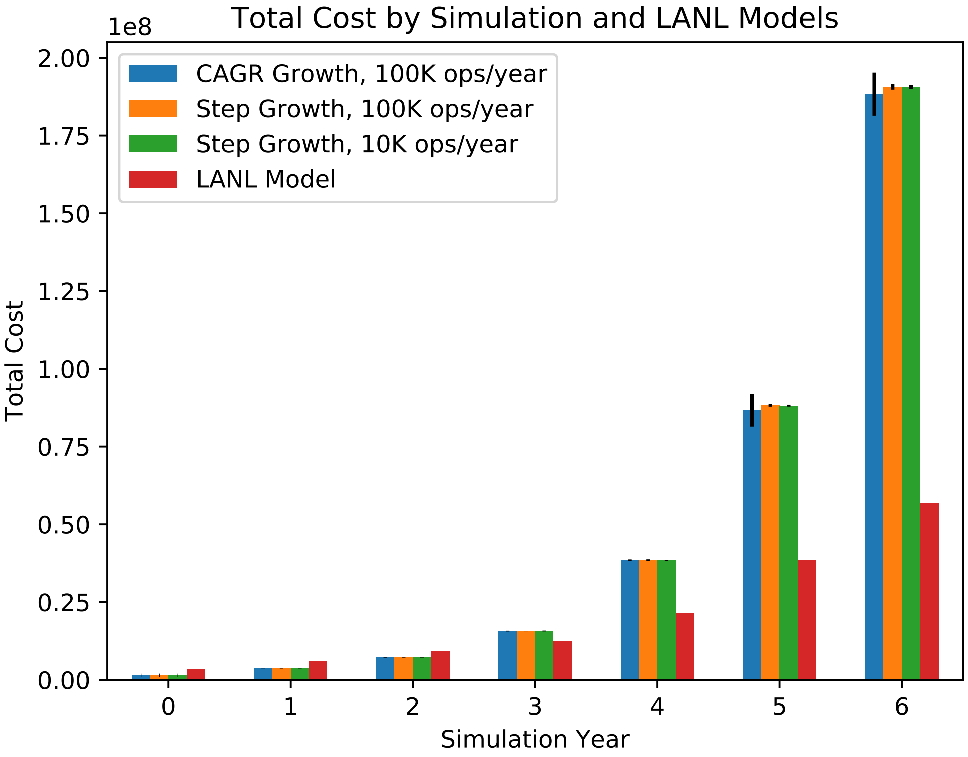

Figure 6.1: We compare the growth of the total cost of an archival system with that from the LANL data set. The models differ in their approaches, and our model includes factors such as electricity cost, device failure and replacement, as well as a different cadence by which generations of new technology become available. The combined effect of the differences between the two models explains their differences.

of the difference between the simulation model and the LANL data arises from the greater number of storage media in the simulation compared with the LANL data. Table 6.4 shows the number of tape cartridges in the simulation for each year. The data in Table 6.4 was drawn from the results for the CAGR growth of media capacity with 100,000 read operations annually. The number of storage devices begins the simulation closely tracking that from the LANL data, but as time passes, the number of storage media grows more quickly than the LANL data so that the number of media and their total cost are approximately double that of the LANL model. The increased number of storage media also increases the demand for tape libraries at a cost of $100,000 each.

The archival system in the simulation closely tracks the capacity demand from Table 6.1; however, the number of media devices needed to store the capacity grows more quickly in in the simulation model than it does in the LANL data. The simulation begins with the first generation of storage media delivering 30 GB as in does in the LANL model; however, the simulation model begins with the first generation of storage technology lasting two full years of simulation time. Since the second generation of tape delivers more than three times the capacity of the first generation, the result of the delay of the new generation is a greater demand for tape media and libraries with all their cost. The delay by one year of the second generation of tape devices within the simulation model also has ripple effects that compound their effects early in the simulation. The third and fourth generations of tape technology also arrive one year later in the simulation compared with the LANL model, and the delay of their added capacity increases the need to purchase more tape media and libraries.

Table 6.4: Number of Tape Media in Simulation

Simulation Year	Number of Media	Total Media Cost
0	16,667	$833,350
1	40,334	$2,021,000
2	57,635	$2,895,752
3	99,349	$5,701,088
4	149,758	$9,101,463
5	269,110	$19,843,206
6	370,119	$28,974,412

6.4 Conclusions

Although the number of tape media in the simulated archival system is greater than that of the LANL model, we find that the simulator produces results that compare meaningfully and explicably with those of other models. The results from our simulation model do not match exactly with the results from other works, yet we would not expect such a match because our simulation model measures different and additional factors than other models and approaches have done in other works. Furthermore, as the difference between our simulation model and the validation data set from LANL exists as a direct result of the differences between their approaches and operating variables, we conclude that our simulation model remains a viable tool for uncovering and exploring differences in storage technologies and the way they are deployed within archival storage systems.

Chapter 7

Conclusion

The ever-increasing demand for archival storage capacity, throughput, and reliability is creating a growing need for new storage technologies that can lower costs while preserving data at scale. Existing storage technologies are struggling to keep pace with demand, and novel storage technologies are needed to meet the future demands for cost-efficient archival storage.

We have described archival storage and defined it in terms of capacity, workload, and duration. We have explained how each storage technology from traditional to novel and prospective presents its own unique characteristics and attendant advantages and challenges relative to archival storage. We have described the functioning of our simulator in detail and presented the parameters that we use to control its operation and generate our results. Next, we presented results on the reliability of archival systems, the cost of serving different workloads using each storage technology, and compared our simulator's output with those of other researchers. Through all our simulations we found that novel and prospective storage technologies like archival glass and synthetic DNA will dominate existing and traditional technologies in most cases. Furthermore, archival glass will help to ensure that the cost of archiving data does not rise to prohibitive levels in the coming years. Synthetic DNA may become a dominant technology for cold archival data storage if it can achieve moderate throughput for writing data. Existing storage technologies like tape, hard disk drives, and solid state disk will be competitive for the next few years, but their limited development roadmaps will cause their economic viability to diminish over time. Traditional storage technologies may nevertheless remain competitive for archival systems that demand the highest degrees of workload intensity of when archival systems must be integrated inside a primary storage system. Each technology that we have considered relies upon the continued work of developers and engineers to refine its technology and accelerate its features to meet the constantly growing demands of archival storage systems.

Bibliography

[1] 45 Drives. The power behind large data storage. http://www.45drives.com/blog/uncategorized/the-power-behind-large-data-storage/, May 2015.

[2] Aalbun. White paper: Innovation and intellectual property rights. https://www.aalbun.com/semiconductor_innovation_ipr_whitepaper-2.0, 2021.

[3] Active Archive Alliance. NERSC exceeds reliability standards with tape-based active archive. https://www.nersc.gov/assets/pubs_presos/AAA-Case-Study-NERSC-FINAL2-6-12.pdf, February 2012.

[4] Ian Adams, Brian Madden, Joel Frank, Mark W. Storer, Ethan L. Miller, and Gene Harano. Usage behavior of a large-scale scientific archive. In *Proceedings of the 2012 International Conference for High Performance Computing, Networking, Storage and Analysis (SC12)*, November 2012.

[5] Ian F. Adams, Mark W. Storer, and Ethan L. Miller. Analysis of workload behavior in scientific and historical long-term data repositories. *ACM Transactions on Storage*, 8(2), 2012.

[6] Jason Adrian and Gordon Goins. Accellerated tape durability testing. https://drive.google.com/file/d/1pPTXKDo7_1tvG_qLC1UmhbxIKsAdAMQj/view, October 2018.

[7] AKCP. How temperature affects IT storage. https://www.akcp.com/blog/how-temperature-affects-it-data-storage/, June 2021.

[8] Patrick Anderson, Richard Black, Ausra Cerkauskaite, Andromachi Chatzieleftheriou, James Clegg, Chris Dainty, Raluca Diaconu, Austin Donnelly, Rokas Drevinskas, Alexander Gaunt, Andreas Georgiou, Ariel Gomez Diaz, Peter G. Kazansky, David Lara, Sergey Legtchenko, Sebastian Nowozin, Aaron Ogus, Douglas Phillips, Ant Rowstron, Masaaki Sakakura, Ioan Stefanovici, Benn Thomsen, and Lei Wang. Glass: A new media for a new era? In *10th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 18)*, pages 1–6, July 2018.

[9] AnySilicon. Semiconductor wafer mask costs. https://anysilicon.com/semiconductor-wafer-mask-costs/, 2022.

[10] ATP Electronics, Incorporated. How temperature affects data retention for SSDs. https://www.atpinc.com/blog/ssd-data-retention-temperature-thermal-throttling, October 2018.

[11] Backblaze, Inc. Hard drive data and stats. https://www.backblaze.com/b2/hard-drive-test-data.html, June 2019.

[12] Shobana Balakrishnan, Richard Black, Austin Donnelly, Paul England, Adam Glass, Dave Harper, Sergey Legtchenko, Aaron Ogus, Eric Peterson, and Ant Rowstron. Pelican: A building block for exascale cold data storage. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI '14)*, October 2014.

[13] Roderick Bauer. SSD 101: How reliable are SSDs? https://www.backblaze.com/blog/how-reliable-are-ssds/, February 2019.

[14] Hugh Bennett. Understanding CD-R and CD-RW. http://www.osta.org/technology/pdf/cdr_cdrw.pdf, January 2003.

[15] Hugh Bennett. Understanding recordable and rewritable DVD. http://www.osta.org/technology/dvdqa/pdf/dvdqa.pdf, April 2004.

[16] R. Bez and A. Pirovano. Overview of non-volatile memory technology: markets, technologies and trends. In Yoshio Nishi, editor, *Advances in Non-volatile Memory and Storage Technology*, pages 1–24. Woodhead Publishing, 2014.

[17] Andrew Binstock. Measuring HDD power usage with newer tech universal adapter. https://www.newertech.com/Static/articles/article_greenercomp_hhdpower.php, October 2007.

[18] Richard Black, Austin Donnelly, Dave Harper, Aaron Ogus, and Ant Rowstron. Feeding the pelican: Using archival hard drives for cold storage racks. In *8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16)*. USENIX Association, June 2016.

[19] Blu-ray Disc Association. Blu-ray disc format. http://www.blu-raydisc.com/Assets/Downloadablefile/White_Paper_General_4th_20150817_clean.pdf, August 2015.

[20] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable economics for a digital planet: Ensuring long-term access to digital information. Technical report, Blue Ribbon Task Force, April 2010.

[21] Business Wire. LTO program outlines generation 8 specifications and extends technology roadmap to 12th generation. `https://www.businesswire.com/news/home/20171017005033/en/LTO-Program-Outlines-Generation-8-Specifications-and-Extends-Technology-Roadmap-to-12th-Generation`, October 2017.

[22] James Byron, Darrell D. E. Long, and Ethan L. Miller. Using simulation to design scalable and cost-efficient archival storage systems. In *Proceedings of the 26th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2018)*, September 2018.

[23] James Byron, Ethan L. Miller, and Darrell D. E. Long. Measuring the cost of reliability in archival systems. In *Proceeding of the Conference on Mass Storage Systems and Technologies (MSST '20)*, October 2020.

[24] Yu Cai, Saugata Ghose, Yixin Luo, Ken Mai, Onur Mutlu, and Erich F. Haratsch. Vulnerabilities in MLC NAND flash memory programming: Experimental analysis, exploits, and mitigation techniques. In *Proceedings of the 23rd International Symposium on High-Performance Computer Architecture (HPCA 2017)*, pages 49–60, February 2017.

[25] Yu Cai, Onur Mutlu, Erich F. Haratsch, and Ken Mai. Program interference in MLC NAND flash memory: Characterization, modeling, and mitigation. In *Proceedings of the 31st International Conference on Computer Design (ICCD)*, pages 123–130, 2013.

[26] M. Calzarossa and G. Serazzi. Workload characterization: A survey. *Proceedings of the IEEE*, 81(8):1136–1150, August 1993.

[27] Maria Carla Calzarossa, Luisa Massari, and Daniele Tessera. Workload characterization: A survey revisited. *ACM Computing Surveys*, 48(3), February 2016.

[28] Rob Carlson. Guesstimating the size of the global array synthesis market. `http://www.synthesis.cc/synthesis/2017/8/guesstimating-the-size-of-the-global-array-synthesis-market`, August 2017.

[29] P. Carns, K. Harms, J. Jenkins, M. Mubarak, R. Ross, and C. Carothers. Impact of data placement on resilience in large-scale object storage systems. In *2016 32nd Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–12, May 2016.

[30] Luis Ceze, Jeff Nivala, and Karin Strauss. Molecular digital data storage using dna. *Nature Reviews Genetics*, May 2019.

[31] Andromachi Chatzieleftheriou, Ioan Stefanovici, Dushyanth Narayanan, Benn Thomsen, and Ant Rowstron. Could cloud storage be disrupted in the next decade? In *12th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 20)*, July 2020.

[32] Fei Chen, Bo Chen, Hongzhe Lin, Yachen Kong, Xin Liu, Xuepeng Zhan, and Jiezhi Chen. Temperature impacts on endurance and read disturbs in charge-trap 3d NAND flash memories. *Micromachines*, 12(10), September 2021.

[33] Leo Chen, Zhonglai Wang, Jing Qiu, Bin Zheng, and Hong-Zhong Huang. Adaptive bathtub hazard rate curve modelling via transformed radial basis functions. 06 2011.

[34] Weida D. Chen, A. Xavier Kohll, Bichlien Nguyen, Julian Koch, Reinhard Heckel, Wendelin J. Stark, Luis Ceze, Karin Strauss, and Robert N. Grass. Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles. *Advanced Functional Materials*, May 2019.

[35] George Crump. The how and why of high-performance primary storage as-a-service. `https://storageswiss.com/2019/10/16/how-and-why-of-high-performance-primary-storage-as-a-service/`, October 2019.

[36] John D. Davis and Steven Swanson. The bleak future of NAND flash memory. In *Proceedings of the 10th USENIX Conference on File and Storage Technologies (FAST '12)*, February 2012.

[37] Yuhui Deng. What is the future of disk drives, death or rebirth? *ACM Computing Surveys*, 43(3), April 2011.

[38] Energy Information Administration. Electric power monthly. `https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a`, November 2021.

[39] Jason Feist. Multi actuator technology: A new performance breakthrough. `https://blog.seagate.com/craftsman-ship/multi-actuator-technology-a-new-performance-breakthrough/`.

[40] flashdba. Understanding flash: Fabrication, shrinkage and the next big thing. `https://flashdba.com/2015/02/17/understanding-flash-fabrication-shrinkage-and-the-next-big-thing/`, February 2015.

[41] Congming Gao, Min Ye, Chun Jason Xue, Youtao Zhang, Liang Shi, Jiwu Shu, and Jun Yang. Reprogramming 3D TLC flash memory based solid state drives. *ACM Transactions on Storage*, 18(1), January 2022.

[42] Gary Grider and Jeff Inman. Unpublished spreadsheet, 2013. Los Alamos National Laboratory.

[43] Akira Goda. Recent progress on 3D NAND flash technologies. *Electronics*, 10(24), 2021.

[44] Phil Goodwin. Tape and cloud: Solving storage problems in the zettabyte era of data. Technical report, IDC, June 2019.

[45] Preeti Gupta, Avani Wildani, Daniel Rosenthal, Ethan L. Miller, Ian Adams, Christina Strong, and Andy Hospodor. An economic perspective of disk vs. flash media in archival storage. In *Proceedings of the 22nd International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '14)*, September 2014.

[46] F. T. Hady, A. Foong, B. Veal, and D. Williams. Platform storage performance with 3D XPoint technology. *Proceedings of the IEEE*, 105(9):1822–1833, 2017.

[47] Jim Handy. 3D NAND: Benefits of charge traps over floating gates. `https://thememoryguy.com/3d-nand-benefits-of-charge-traps-over-floating-gates/`, November 2013.

[48] Jim Handy. 3D NAND: Making a vertical string. `https://thememoryguy.com/3d-nand-making-a-vertical-string/`, November 2013.

[49] T. Heer. L2 milestone: High performance next-gen tape archive deployed to classified computing environment at LLNL. Technical report, Lawrence Livermore National Laboratory, September 2020.

[50] Hewlett Packard Enterprise. HPE LTO ultrium cartridges. `https://psnow.ext.hpe.com/doc/PSN34648USEN.pdf`, 2018.

[51] Hewlett Packard Enterprise. HPE LTO-8 ultrium 30TB RW data cartridge data sheet. `https://www.hpe.com/psnow/doc/PSN1010419339CAEN.pdf`, 2022.

[52] HGST. Ultrastar HS14. `https://www.hgst.com/sites/default/files/resources/Ultrastar-Hs14-DS.pdf`, 2017.

[53] James P Hughes. Economics of information storage: The value in storing the long tail. In *2019 35th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 185–192. IEEE, 2019.

[54] Avery Hurt. Is holographic data storage the next big thing? `https://www.discovermagazine.com/technology/is-holographic-data-storage-the-next-big-thing`, January 2022.

[55] Chioma Ibeakanma. What is DNA data storage? is it the future of storage? https://www.makeuseof.com/what-is-dna-data-storage/, April 2022.

[56] Ilias Iliadis, Jens Jelitto, Yusik Kim, Slavisa Sarafijanovic, and Vinodh Venkatesan. ExaPlan: Efficient queueing-based data placement, provisioning, and load balancing for large tiered storage systems. *ACM Transactions on Storage*, 13(2):17–41, May 2017.

[57] Jeff Inman, Gary Grider, and Hsing Bung Chen. Cost of tape versus disk for archival storage. In *2014 IEEE 7th International Conference on Cloud Computing (CLOUD '14)*, pages 208–215, July 2014.

[58] International Business Machines, Inc. RAMAC: the first magnetic hard disk. https://www.ibm.com/ibm/history/ibm100/us/en/icons/ramac/.

[59] International Business Machines, Incorporated. Environmental and shipping specifications for LTO tape cartridges. https://www.ibm.com/docs/en/ts3500-tape-library?topic=media-environmental-shipping-specifications-lto-tape-cartridges, March 2021.

[60] Latchesar Ionkov and Bradley Settlemyer. DNA: The ultimate data-storage solution. *Scientific American*, May 2021.

[61] David W. Jensen and Daniel A. Reed. File archive activity in a supercomputer environment. Technical Report UIUCDCS-R-91-1672, University of Illinois at Urbana-Champaign, April 1991.

[62] David W. Jensen and Daniel A. Reed. File archive activity in a supercomputing environment. In *Proceedings of the 7th International Conference on Supercomputing*, ICS '93, page 387âĂŞ396, New York, NY, USA, 1993. Association for Computing Machinery.

[63] Y. Jia and F. Chen. From flash to 3D XPoint: Performance bottlenecks and potentials in RocksDB with storage evolution. In *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 192–201, 2020.

[64] Theodore Johnson and Ethan Miller. Benchmarking tape system performance. In *Proceedings of the 6th Goddard Conference on Mass Storage Systems and Technologies / 15th IEEE Symposium on Mass Storage Systems*, February 1998.

[65] Andy Klein. Hard drive stats for Q3 2016: Less is more. https://www.backblaze.com/blog/hard-drive-failure-rates-q3-2016/, November 2016.

[66] Andy Klein. Hard drive stats for Q3 2018: Less is more. `https://www.backblaze.com/blog/2018-hard-drive-failure-rates/`, October 2018.

[67] David M. Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, and Heidi Zimmer. Introduction to statistics: Online edition. `https://onlinestatbook.com/Online_Statistics_Education.pdf`, 2003.

[68] DongJin Lee, Michael O'Sullivan, Cameron Walker, and Monique MacKenzie. Robust benchmarking for archival storage tiers. In *Proceedings of the 6th Parallel Data Storage Workshop (PDSW '11)*, November 2011.

[69] Kyunghwan Lee, Myounggon Kang, Seongjun Seo, Dong Hua Li, Jungki Kim, and Hyungcheol Shin. Analysis of failure mechanisms and extraction of activation energies (e_a) in 21-nm NAND flash cells. *IEEE Electron Device Letters*, 34(1):48–50, 2013.

[70] Hai Li. Storage physics and noise mechanism in heat-assisted magnetic recording. Technical report, Carnegie Mellon University, September 2016.

[71] Wenhao Li, Yun Yang, and Dong Yuan. Generic data reliability model in the cloud. In Wenhao Li, Yun Yang, and Dong Yuan, editors, *Reliability Assurance of Big Data in the Cloud*, pages 31–36. Morgan Kaufmann, 2015.

[72] Wenhao Li, Yun Yang, and Dong Yuan. Literature review. In Wenhao Li, Yun Yang, and Dong Yuan, editors, *Reliability Assurance of Big Data in the Cloud*, pages 9–17. Morgan Kaufmann, 2015.

[73] Wenhao Li, Yun Yang, and Dong Yuan. Minimum replication for meeting the data reliability requirement. In Wenhao Li, Yun Yang, and Dong Yuan, editors, *Reliability Assurance of Big Data in the Cloud*, pages 37–43. Morgan Kaufmann, 2015.

[74] Wenhao Li, Yun Yang, and Dong Yuan. Motivating example and problem analysis. In Wenhao Li, Yun Yang, and Dong Yuan, editors, *Reliability Assurance of Big Data in the Cloud*, pages 19–29. Morgan Kaufmann, Boston, 2015.

[75] Yan Li, Darrell D. E. Long, and Ethan L. Miller. Understanding data survivability in archival storage systems. In *Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR '12)*, June 2012.

[76] LTO Consortium. LTO program releases specifications for upcoming generation 9. `https://www.lto.org/2020/09/the-lto-program-releases-specifications-for-upcoming-generation-9/`, September 2020.

[77] LTO Consortium. LTO generation compatibility. https://www.lto.org/lto-generation-compatibility/, 2022.

[78] LTO Consortium. Roadmap. https://www.lto.org/roadmap/, 2022.

[79] Gough Lui. Hard drive performance over the years. https://goughlui.com/the-hard-disk-corner/hard-drive-performance-over-the-years/, 2022.

[80] Fabio Margaglia and Andre Brinkmann. Improving MLC flash performance and endurance with extended P/E cycles. In *Proceedings of the 31st IEEE Conference on Mass Storage Systems and Technologies*, pages 1–12, May 2015.

[81] Chris Mellor. QLC flash is tricky stuff to make and use, so here's a primer. https://www.theregister.co.uk/2016/07/28/qlc_flash_primer/, July 2016.

[82] Chris Mellor. WD shoots out 96-layer embedded flash chips. https://www.theregister.co.uk/2018/10/18/wds_96layer_embedded_flash_chips/, October 2018.

[83] Chris Mellor. LTO-8 tape media patent lawsuit cripples supply as Sony and Fujifilm face off in court. https://www.theregister.com/2019/05/31/lto_patent_case_hits_lto8_supply/, May 2019.

[84] Chris Mellor. Sony and Fujifilm settle LTO-8 tape media patent dispute. https://blocksandfiles.com/2019/08/06/lto-8-tape-media-manufacturing-block-mysteriously-blown-away/, August 2019.

[85] Chris Mellor. Seeing the light: Folio Photonics hopes to crack the optical archive disk market. https://blocksandfiles.com/2021/08/31/seeing-the-archive-light-folio-photonics-hopes-to-crack-the-optical-archive-disk-market-wide-open/, August 2021.

[86] Chris Mellor. Western Digital: The flash roadmap. https://blocksandfiles.com/2022/05/12/western-digital-flash-roadmap/, May 2022.

[87] Daniel A. Menasce. CS 672: Workload characterization. https://cs.gmu.edu/~menasce/cs672/slides/CS672-wkldchar.pdf, 1999.

[88] Justin Meza, Qiang Wu, Sanjeev Kumar, and Onur Mutlu. A large-scale study of flash memory failures in the field. In *Proceedings of the 2015 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, June 2015.

[89] R. Micheloni and L. Crippa. Multi-bit NAND flash memories for ultra high density storage devices. In Yoshio Nishi, editor, *Advances in Non-volatile Memory and Storage Technology*, pages 75–119. Woodhead Publishing, 2014.

[90] Ethan Miller and Randy Katz. An analysis of file migration in a Unix supercomputing environment. In *Proceedings of the Winter 1993 USENIX Technical Conference*, pages 421–433, January 1993.

[91] Asit K. Mishra, Joseph L. Hellerstein, Walfredo Cirne, and Chita R. Das. Towards characterizing cloud backend workloads: Insights from google compute clusters. In *Proceedings of the 2010 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, volume 37, page 34âĂŞ41. Association for Computing Machinery, March 2010.

[92] Chuang-Hue Moh. TimeLine: A high performance archive for a distributed object store. In *First Symposium on Networked Systems Design and Implementation (NSDI 04)*. USENIX Association, March 2004.

[93] Samuel K. Moore. Micron is first to deliver 3D flash chips with more than 200 layers. `https://spectrum.ieee.org/micron-is-first-to-deliver-3d-flash-chips-with-more-than-200-layers`, July 2022.

[94] S. Morup, M.F. Hansen, and C. Frandsen. Magnetic nanoparticles. In David L. Andrews, Gregory D. Scholes, and Gary P. Wiederrecht, editors, *Comprehensive Nanoscience and Technology*, pages 437–491. Academic Press, Amsterdam, 2011.

[95] MP Tapes, Incorporated. Reliability of magnetic data tape. `https://mptapes.com/Reliability/reliability.html`, November 2017.

[96] National Instruments Corporation. Effects of temperature on SSD endurance on LabVIEW real-time systems. `https://www.ni.com/en-us/support/documentation/supplemental/18/effects-of-temperature-on-ssd-endurance.html`, February 2022.

[97] Sharon Newman, Ashley Stephenson, Max Willsey, Bichlien Nguyen, Christopher N. Takahashi, Karin Strauss, and Luis Ceze. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nature Communications*, 9, April 2019. (2019)10:1706.

[98] O. Ozatay, P. G. Mather, J.-U. Thiele, T. Hauet, and P. M. Braganca. Spin-based data storage. In David L. Andrews, Gregory D. Scholes, and Gary P. Wiederrecht, editors, *Comprehensive Nanoscience and Technology*, pages 561–614. Academic Press, Amsterdam, 2011.

[99] Panasonic Corporation. Panasonic and Facebook develop optical disc-based data archive system for data centers. `https://phys.org/news/2016-01-panasonic-facebook-optical-disc-based-archive.html`, January 2016.

[100] Panasonic Corporation. Optical data archive. `ftp://ftp.panasonic.com/datastorage/intelligentarchive_referencearchitecture_brochure.pdf`, 2017.

[101] J. Pâris, S. J. T. Schwarz, and D. D. E. Long. Improving disk array reliability through faster repairs (extended abstract). In *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*, pages 1–2, December 2016.

[102] Jehan-François Pâris and Darrell D. E. Long. Using device diversity to protect data against batch-correlated disk failures. In *Proceedings of the Second ACM Workshop on Storage Security and Survivability*, pages 47–52, New York, NY, USA, 2006. ACM.

[103] Joseph Pasquale, Barbara Bittel, and Daniel Kraiman. A static and dynamic workload characterization study of the San Diego supercomputer center Cray X-MP. In *Proceedings of the 1991 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '91, page 218âĂŞ219. Association for Computing Machinery, 1991.

[104] David A. Patterson, Garth Gibson, and Randy 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.

[105] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz AndrÃľ Barroso. Failure trends in a large disk drive population. In *5th USENIX Conference on File and Storage Technologies (FAST 2007)*, pages 17–29, 2007.

[106] Davide Resnati, Akira Goda, Gianluca Nicosia, Carmine Miccoli, Alessandro S. Spinelli, and Christian Monzio Compagnoni. Temperature effects in NAND flash memories: A comparison between 2-D and 3-D arrays. *IEEE Electron Device Letters*, 38(4):461–464, February 2017.

[107] Drew Robb. Is optical disc an illusion? `https://www.enterprisestorageforum.com/storage-technology/is-optical-disc-an-illusion.html`, January 2016.

[108] Les Robertson. Data storage technologies for LHC. Technical report, CERN, 2000.

[109] David Rosenthal. Economic models of long-term storage. https://blog.dshr.org/2019/02/economic-models-of-long-term-storage.html, February 2019.

[110] David S.H. Rosenthal, Daniel Rosenthal, Ethan L. Miller, Ian Adams, Mark W. Storer, and Erez Zadok. The economics of long-term digital storage. In *The Memory of the World in the Digital Age: Digitization and Preservation*, September 2012.

[111] Samsung Electronics. Samsung V-NAND SSD: 860 EVO. https://www.samsung.com/semiconductor/global.semi.static/Samsung_SSD_860_EVO_Data_Sheet_Rev1.pdf, December 2017.

[112] Felipe Garcia Sanchez. Modeling of field and thermal magnetization reversal in nanostructured magnetic materials. Technical report, Universidad AutÃşnoma de Madrid, November 2007.

[113] Bianca Schroeder, Raghav Lagisetty, and Arif Merchant. Flash reliability in production: The expected and the unexpected. In *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST '16)*, pages 67–80, February 2016.

[114] Seagate Technology. Choosing high performance storage isn't just about RPM. https://www.seagate.com/tech-insights/choosing-high-performance-storage-is-not-about-rpm-anymore-master-ti/.

[115] Seagate Technology. Archive HDD data sheet. https://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/archive-hdd-dS1834-3-1411us.pdf, 2014.

[116] Serve the Home. RAID reliability calculator. https://www.servethehome.com/raid-calculator/raid-reliability-calculator-simple-mttdl-model/.

[117] Anuj R. Shah, Joshua N. Adkins, Douglas J. Baxter, William R. Cannon, Daniel G. Chavarria-Miranda, Sutanay Choudhury, Ian Gorton, Deborah K. Gracio, Todd D. Halter, Navdeep D. Jaitly, John R. Johnson, Richard T. Kouzes, Matthew C. Macduff, Andres Marquez, Matthew E. Monroe, Christopher S. Oehmen, William A. Pike, Chad Scherrer, Oreste Villa, Bobbie-Jo Webb-Robertson, Paul D. Whitney, and Nino Zuljevic. Chapter 1 - applications in data-intensive computing. In *Advances in Computers*, volume 79, pages 1–70. Elsevier, 2010.

[118] Robert Sheldon. WORM (write once, read many). https://www.techtarget.com/searchstorage/definition/WORM-write-once-read-many, January 2022.

[119] R. Shirota. Developments in 3D-NAND flash technology. In Yoshio Nishi, editor, *Advances in Non-volatile Memory and Storage Technology*, pages 27–74. Woodhead Publishing, 2014.

[120] S. R. Shishira, A. Kandasamy, and K. Chandrasekaran. Workload characterization: Survey of current approaches and research challenges. In *Proceedings of the 7th International Conference on Computer and Communication Technology*, ICCCT-2017, page 151âĂŞ156. Association for Computing Machinery, 2017.

[121] Hubbert Smith. Using QLC for cold storage is a fool's errand. `https://blocksandfiles.com/2019/09/27/using-qlc-for-cold-storage-is-a-fools-errand/`, September 2019.

[122] Koji Sonoda. Flying instability due to organic compounds in hard disk drive. *Advances in Tribology*, 2012, December 2012.

[123] Sony Corporation. Optical disc archive, generation 2: White paper. `http://assets.pro.sony.eu/Web/ngp/pdf/optical-disc-archive-generation-two.pdf`, April 2016.

[124] Sony Corporation and Panasonic Corporation. White paper: Archival disc technology. `https://panasonic.net/cns/archiver/pdf/E_WhitePaper_ArchivalDisc_Ver100.pdf`, July 2015.

[125] Mark W. Storer, Kevin M. Greenan, Ethan L. Miller, and Kaladhar Voruganti. Pergamum: Replacing tape with energy efficient, reliable, disk-based archival storage. In *Proceedings of the 6th USENIX Conference on File and Storage Technologies*, FAST'08. USENIX Association, 2008.

[126] Christopher N. Takahashi, Bichlien Nguyen, Karin Strauss, and Luis Ceze. Demonstration of end-to-end automation of DNA data storage. *Nature Scientific Reports*, 9, March 2019. Article number: 4998 (2019).

[127] Yoshiki Takai, Mamoru Fukuchi, Reika Kinoshita, Chihiro Matsui, and Ken Takeuchi. Analysis on heterogeneous SSD configuration with quadruple-level cell (QLC) NAND flash memory. In *Proceedings of the 11th International Memory Workshop (IMW)*, pages 1–4, May 2019.

[128] The International Disk Drive Equipment and Materials Association. ASTC technology roadmap. `http://idema.org/wp-content/plugins/download-monitor/download.php?id=2456`, 2016.

[129] Keri Troutman. NERSC tape archives make the move to Berkeley lab's Shyh Wang Hall. `https://www.nersc.gov/news-publications/nersc-`

news/nersc-center-news/2019/nersc-tape-archives-make-the-move-to-berkeley-labs-shyh-wang-hall/, feb 2019.

[130] Carmen Valache. Blast from the past: Retrieving your data from old storage media. https://interestingengineering.com/diy/blast-from-the-past-retrieving-your-data-from-old-storage-media, October 2019.

[131] V. Venkatesan, I. Iliadis, C. Fragouli, and R. Urbanke. Reliability of clustered vs. declustered replica placement in data storage systems. In *2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 307–317, July 2011.

[132] Versity Software, Inc. Designing a high performance tape archive. https://www.versity.com/designing-a-high-performance-tape-archive/, July 2018.

[133] Chip Walter. Kryder's law. https://www.scientificamerican.com/article/kryders-law/, August 2005.

[134] Joseph C. Watkins. An introduction to the science of statistics: From theory to implementation. https://www.math.arizona.edu/~jwatkins/statbook.pdf.

[135] Charles M. Weber, C. Neil Berglund, and Patricia Gabella. Mask cost and profitability in photomask manufacturing: An empirical analysis. *IEEE Transactions on Semiconductor Manufacturing*, 19(4):465–474, 2006.

[136] A. Wildani and I. F. Adams. A case for rigorous workload classification. In *2015 IEEE 23rd International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 146–149, 2015.

[137] Wintelguy. RAID reliability calculator. https://wintelguy.com/raidmttdl.pl.

[138] Gala Yadgar and Moshe Gabel. Avoiding the streetlight effect: I/O workload analysis with SSDs in mind. In *8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16)*, Denver, CO, June 2016. USENIX Association.

[139] Zihui Yan and Cong Liang. New levenshtein-marker code for DNA-based data storage capable of correcting multiple edit errors. September 2021.

[140] Jinfeng Yang, Bingzhe Li, and David J. Lilja. Exploring performance characteristics of the optane 3D Xpoint storage technology. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 5(1), February 2020.

[141] David Yu, Guangwei Che, Tim Chou, and Ognian Novakov. Best practices in accessing tape-resident data in HPSS. *The European Physical Journal Web Conferences*, 214, 2019.

[142] J. Zhang, P. Li, B. Liu, T. G. Marbach, X. Liu, and G. Wang. Performance analysis of 3D XPoint SSDs in virtualized and non-virtualized environments. In *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1–10, 2018.

[143] Victor Zhirnov, Reza M. Zadegan, Gurtej S. Sandhu, George M. Church, and William L. Hughes. Nucleic acid memory. *Nature materials*, 15, February 2016.

[144] Huijun Zhu, Peng Gu, and Jun Wang. Shifted declustering: a placement-ideal layout scheme for multi-way replication storage architecture. In Pin Zhou, editor, *Proceedings of the 22nd Annual International Conference on Supercomputing, ICS 2008, Island of Kos, Greece, June 7-12, 2008*, pages 134–144. ACM, June 2008.