Glass: A New Media for a New Era?

Patrick Anderson,[1] Richard Black,[1] Aušra Čerkauskaitė,[2] Andromachi Chatzieleftheriou,[1] James Clegg,[1] Chris Dainty,[1] Raluca Diaconu,[1] Austin Donnelly,[1] Rokas Drevinskas,[1] Alexander L. Gaunt,[1] Andreas Georgiou,[1] Ariel Gomez Diaz,[1] Peter G. Kazansky,[2] David Lara,[1] Sergey Legtchenko,[1] Sebastian Nowozin,[1] Aaron Ogus,[1] Douglas Phillips,[1] Antony Rowstron,[1] Masaaki Sakakura,[2] Ioan Stefanovici,[1] Benn Thomsen,[1] Lei Wang,[2] Hugh Williams,[1] and Mengyang Yang[1]

[1]Microsoft Research
[2]Optoelectronics Research Centre, University of Southampton

Abstract

In the foreseeable future, major cloud vendors will need to store multiple zettabytes of data in their cloud storage infrastructure. Like all storage, cloud storage systems need to trade performance for cost, and they currently achieve this by using storage tiers backed by different storage media. The ultimate goal for cloud storage would be to provide zero-cost storage with low access latencies and high throughput. But all the storage media being deployed in the cloud today were created before the cloud existed, and were designed to support many usage scenarios. With cloud storage, the focus is on cost, and storage needs to be designed to be right-provisionable. The limits of what is possible with existing storage technologies are being reached, and a new clean-slate approach is needed for cloud storage. Hence, the time is right to seek out a new physical media to underpin novel storage systems designed exclusively to support the cloud. In Project Silica, Microsoft Research and Southampton University, are exploring whether quartz glass could be the future media for mass storage in the cloud. In this paper, we describe the basis for the technology, and discuss conventional assumptions about storage that we are challenging in Project Silica.

1 Introduction

Over the past decade, much of the world's data has moved into the cloud. The demand for long-term cloud storage has reached unprecedented levels, and it's expected that by the beginning of the next decade, zettabytes of data will be stored in the cloud [9]. To cope with the ever-increasing storage demand, cloud providers rely on many storage technologies including non-volatile memory (NVM), flash, hard disk drives (HDDs), magnetic tape, and even optical discs; all with vastly different characteristics in terms of: cost, latency, throughput, storage density, failure characteristics, and media lifetime.

The first observation, is that all these storage technologies were designed before the existence of the cloud. The second observation is that, because of the different properties of the storage technologies, no one storage dominates. If you had to pick a single technology, flash would be it, as it offers acceptable durability, high throughput, and low latency storage. A flash-provisioned service can handle all current workloads in the cloud and could offer a single storage Service Level Agreement (SLA). However, the cost of flash makes it prohibitive for storing anything other than hot data. This forces cloud providers to offer segregated tiers that service different segments of their customers' data, each with a different SLA, as the pricing and performance characteristics of each tier is dictated by the properties of the storage technology underpinning it. The third observation is that HDDs are used as the predominant mass storage technology in the cloud. The majority of bytes stored in the cloud are stored on HDDs as they offer much lower \$/GB costs than NVM and flash, leading to lower total costs over the lifetime of the data. However, it is widely acknowledged that HDDs have a significant number of problems [3].

These observations motivate us to consider what would be the best storage media for high-volume long-lived cool to cold data, that currently resides on "capacity-oriented" storage technologies (i.e., archival HDDs and magnetic tapes), that would offer better access latencies at lower cost. The cheapest option today is probably magnetic tape. But even under low load, a tape library has an average-case response time on the order of minutes to hours. Nobody aspires to wait hours to access their data, and indeed this is orders of magnitude off from the 10 second limit that will keep a user's attention [12].

A new storage technology is needed for the cloud, and we are currently building a clean-slate storage system designed from scratch to service the cloud. The media that this storage system will use is glass. In this paper, we will describe the breakthroughs that enabled storing data in glass, and the opportunities that designing end-to-end down to the materials level afford us to change how we think of and build storage systems. As we understand how to build a storage system around glass, there are three fundamental problems and limitations in today's storage systems we would like to address: *entanglement*, *refresh cycle* and *constrained workloads*.

Entanglement: Today's systems are unable to scale their write and read throughputs independently from each other, and separately from the cost of storing the data. Scaling performance is done by deploying more devices, which include read and write capabilities, as well as the physical media. The aggregate write throughput should support just the data being ingested by the system. The aggregate read throughput should support the read demand, which is a function of the total volume of data stored in the system.

Refresh cycle: The existing technologies suffer from requiring a refresh cycle. Over time, the media becomes less reliable; magnetic media are prone to whole-device [15, 14] and partial failures, through bit rot that manifests as latent sector errors [1]. Long-lived data must get copied to new storage media every few years. The decision for when to do this is a function of the annualized failure rate (AFR) and partial failures rate, which increase with time in use [1]. At an extreme, in cold storage, a significant amount of time is spent copying data into and out of a rack. With a Pelican [2] the first month and last month are spent copying data in or out of the rack.

Constrained workloads: Tiering is the norm because across different storage technologies, the performance per $ differs. Users pick the cheapest tier that meets their minimum performance requirements, and then ensure their workload is optimized to suit the performance profile of the tier. This constrains applications: they must partition their data between different tiers based on access characteristics and transfer the data between tiers if these requirements change. We need a mass storage system that leaves the data in situ and is low cost, with acceptable performance across a wide range of access characteristics, to free the users from having to make these choices.

A fundamental consequence of entanglement and the refresh cycle is that the cost of storing data scales with its lifetime rather than the operations performed on it.

We start by describing the pitfalls of current media, then describe the breakthroughs on how to store data in glass. One of the challenges is to accurately read the data stored in the glass, and we describe how we are using machine learning to help address the challenges. We are early in the process of taming this breakthrough, and do not yet have a full end-to-end working system, but we conclude by talking about some of the higher-level system design principles and factors influencing our design.

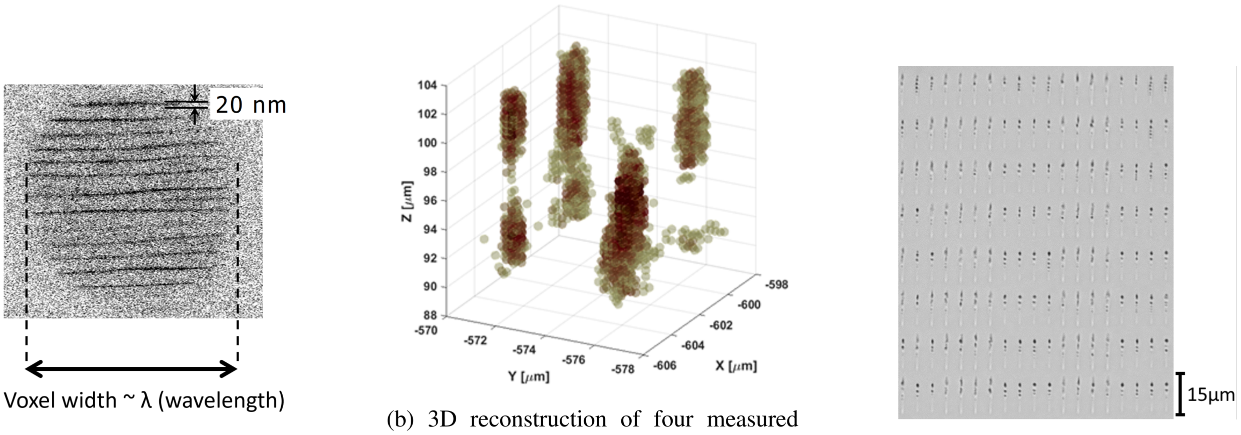
2 Current Storage Technologies

Hard disk drive For over 60 years areal density increases drove performance increases for HDDs. Areal density increases on PMR HDDs have reached their limits. Most recently, the largest driver of capacity increases has been increasing the number of platters, but this is not sustainable due to physical constraints of the 3.5″ HDD casing. SMR [19] delivers marginal increases at the cost of removing random writes. HAMR [16] is not yet commercially viable, and MAMR [5, 23, 22] may become available next year. Further, the bottleneck resource for many workloads is the IOPS/GB the drive can sustain. The IOPS/GB limit can leave capacity stranded, effectively unusable without careful IO management across disks. Proposals to increase the number of heads per drive [3] will help, but will only scale performance to the number of platters in the system, yielding at most an 8-fold performance improvement.

Magnetic tape Tape provides lower $/GB costs than HDDs, but only for cold archival workloads, with an average access latency on the order of minutes to hours [6]. Tape libraries, robots and tape drives must be carefully managed and require frequent maintenance. They require environmentally-controlled environments to operate and need complex total system serviceability (rather than component-level). Without the right environment, tapes can degrade.

Optical storage Optical disc technologies, like Blu-ray, store and read data based on changes made to a thin film made from organic or inorganic materials. A laser encodes data by creating alternating marks and spaces on the surface of the thin film. High capacity optical discs use multiple thin film layers sandwiched together. The limit on density is that each layer needs to have sufficient transmittance to allow the laser beam to write and read the deeper layers. Currently, for a single disc this is about 4-8 layers, which limits the maximum capacity of a disc to less than 1 TB. As the disc is composed of layers of polycarbonate and thin films glued together, the media has a long, but limited lifetime.

Holographic Data Storage In holographic storage [7, 4], writing consists of interfering two laser beams (think of one beam as an address line and the other encodes a data page to be written) inside a block of photorefractive material. Excited electrons transition to higher energy states and move between different types of atoms in the compound. This produces a stable, but reversible change in the refractive properties of the material. Reading consists of using the address line beam only to retrieve the data page from the material. Data is written and read in multi-megabyte pages and can be erased and overwritten. However, the $/GB that holographic storage provides is not competitive with other storage technologies [4], so it has not been deployed at volume commercially.



(a) Single voxel (top view)

(b) 3D reconstruction of four measured voxels

(c) Side view (8 layers)

Figure 1: Voxels in fused silica

3 Storing Data in Glass

A recent breakthrough at the University of Southampton [21, 17, 20] has made it possible to store data in fused silica (i.e., quartz glass). When the beam from a femtosecond laser is focused inside a block of fused silica, a permanent small 3D nanostructure (which we will call a "voxel") forms in the silica. In contrast to holographic storage, writing a voxel in glass involves inducing a permanent, long-term stable change to the physical structure of the fused silica material. Viewed from the top, the voxel has a grating structure (a nanograting) and is circular, with a diameter of approximately the wavelength of the light used to create it, and has a depth of a few microns into the glass. In contrast to conventional optical disc storage, many layers of voxels (i.e., over 100) can be written in glass, as the transmittance of fused silica is much larger than that of opaque thin films used in conventional optical discs, allowing light to penetrate much deeper into the material, for both writing and reading.

Each voxel has a property called *form birefringence*, whereby the nanostructure has different physical properties than the surrounding silica material. The voxel exhibits a different refractive index for light with a different polarization (i.e., light that has its electric field oscillating in a particular direction). As a result, when polarized light interacts with the voxel, a shift of several nanometers between the components of its electric field is introduced. The range of this shift is known as the voxel's "retardance". This shift also induces a change in the polarization angle of incoming light. Retardance and angle change can be used to encode multiple bits per voxel. When writing into glass, modulating the polarization of the laser beam, energy, and the number of pulses makes it possible to deterministically affect these two properties of the created voxel. This allows specific values to be encoded into each voxel. Reading the data stored amounts to measuring these two properties of the voxels.

Figure 1a shows a zoomed-in image of a voxel from the top taken with a scanning electron microscope. In the figure you can see the circular voxel, and the nanograting structure. Figure 1b shows an experimental reconstruction of the 3D structure of four voxels written with a 1030 nanometer wavelength laser. This data was collected using a confocal imaging technique [13] that samples the glass every 0.2 microns in each of the X,Y and Z dimensions. In this example, the width of the nanostructure is around 1 micron, and the depth is around 10 microns. The figure clearly shows the three-dimensional structure of the voxels, and the variation between each voxel in the Z dimension shows some of the challenges that still exist in controlling the write process.

By focusing the laser beam at different positions across an XY plane, we can write voxels side-by-side to form a 2D layer. By changing the focus depth of the laser beam, we can write many layers across the depth of the silica block. Figure 1c shows a side view of 8 layers of voxels.

Once written, a voxel remains stable, and retains its birefringent properties for the lifetime of the fused silica (i.e. many hundreds of years), withstanding temperatures of over 1000°C [21] without any negative impact. As such, quartz glass provides very different properties from existing storage media. Storing data in glass is more akin to stone etchings and writing on paper than to current magnetic storage technologies. It is even electromagnetic field (EMF) proof as a storage media. Magnetic storage technologies are notoriously prone to bit rot, require multiple forms of error detection and correction, and are ultimately lifetime-limited, requiring wholesale copying of data every few years to new media to prevent data corruption and loss. Glass provides a write-once-read-many (WORM) media; the cost of fused silica is low (on the order of a few cents for a piece the size of a DVD), and it offers long-term stability and durability.

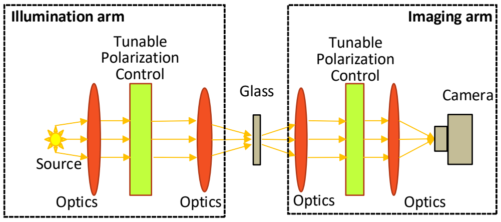
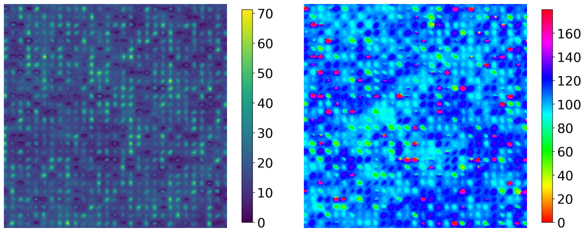


Figure 2: Read head hardware



(a) Retardance (in nm) (b) Angle change (in deg.)

Figure 3: Computed measures of birefringence

There are multiple challenges to using glass as a media. The most obvious is building a storage system that is able to exploit the glass media properties, particularly the lifetime. Further, different technologies are used for writing and reading. The write path uses a femtosecond pulse laser to generate the pulses necessary to create the voxel nanostructures in the glass. This is a different type of laser than diode lasers used in DVD and Blu-Ray drives, and due to its form factor, power and cooling needs, it will be the size of a 2U server. The voxels can be read from the glass using microscopy (e.g. a camera along with optical components akin to a microscope). It should be noted that the multiple voxels can be imaged concurrently, provided they are on the same layer. The methods used to decode the data stored by the voxels will be discussed in Section 4.

4 Voxel Reading

To make glass into a usable storage media we need to efficiently retrieve the stored data. This has two challenges; *(i)* encode as many bits as possible per voxel, and *(ii)* read voxels through 100s of layers.

Recall that data is encoded in glass based on the interaction between a polarized light wave and the birefringent voxels. A voxel has two physical properties that affect a polarized light wave: retardance, and change in polarization angle. Fortunately, measuring and quantifying birefringence in materials is well-studied. Many crystals, plastic materials (particularly under stress), and biological materials exhibit birefringence.

Figure 2 diagrammatically shows the key components required to measure birefringence in glass. It has a light source in an illumination arm, the glass sample, and then an imaging arm with a camera. Both arms have tunable polarization control (e.g., liquid crystal polarizers), which allows you to arbitrarily change the polarization of light passing through each arm. To read birefringence, you need to probe the glass with different kinds of polarized light and measure the observed changes in polarization induced by the voxels.

We sequentially take a set of images of the same field of view. The illumination arm creates a beam of light polarized to one angle. As light passes through the voxels, they change the polarization of the beam. The tunable polarization control and camera in the imaging arm are used to detect those changes. A different set of configurations is used for each image. Conceptually, this is analogous to measuring the projections of a vector onto the bases of the vector space it occupies. To read different layers, the optics focus on a different depth in the glass.

Traditionally, the set of images are combined and processed to determine the retardance and polarization angle change across voxels. Figure 3 shows the two output images for a widely used *Four-Frame Algorithm* [18]. Figure 3(a) shows retardance and (b) shows polarization angle change for a field of view that contains $\sim$ 840 voxels. This algorithm requires four measurement images, plus an additional four background images that quantify any baseline retardance in the block of silica, independent of voxels. Decoding the data value of each voxel then requires sampling the voxel positions in the two output images and thresholding the sampled values into fixed ranges corresponding to multi-bit values.

While there is a deep understanding of how to measure birefringence, there are multiple challenges. The accuracy with which we measure it impacts the performance of the system. The traditional approaches suffer as the layers outside the layer of focus scatter light, which manifests as noise when performing a read. This noise increases significantly as the lateral spacing between the voxels is decreased and as the number of layers is increased. This causes a significant decrease in decoding accuracy. As seen in Figure 1b, individual voxel shapes vary, further adding complexity.

This all impacts decoding performance. We observe that we need an end-to-end approach. For each voxel we want *just* to determine the multi-bit value encoded in it. We are not interested in the absolute magnitude of birefringence per se. These traditional approaches are useful for applications like quality control, where the absolute magnitude of birefringence is directly tied to the amount of physical stress observed by a part (for example).

We exploit this observation by creating a prototype of a decoder optimized for glass media. It uses an end-to-end deep learning approach. The key insight is to replace all intermediate processing steps (i.e., computing

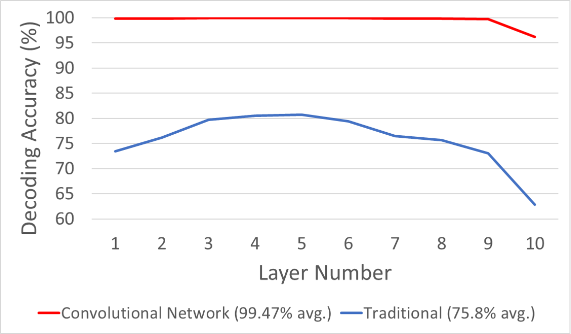


Figure 4: Decoding accuracy

retardance and angle change, thresholding) with a neural network that takes the *raw measurement images* as inputs and produces the *voxel values* as outputs.

This approach makes no preconceived assumptions about which parts of the signal are important. The neural network learns the best internal representation of the voxels necessary to maximize decoding accuracy. The approach we take is inspired by the success of deep learning on many signal processing problems in recent years: images [11], video [10], speech [8], etc.

Using machine learning shifts the complexity and logic necessary to deal with potential variabilities and noise from an online problem (when reading the data) into an offline problem, when training the decoder. Coping with variability simply requires enlarging the training set to include examples of it. The neural network also learns the noise patterns, so does not require any background images. This has the added benefit of reducing the amount of data that needs capturing at runtime.

Figure 4 shows decoding accuracy as a function of layer depth for voxels written at two micron lateral separation, over 10 layers. It shows results for our prototype decoder (using a convolutional neural network) and the traditional Four Frame algorithm. We believe that there is the opportunity to potentially optimize the traditional approach further. However, currently, the prototype decoder achieves an accuracy of 99.47% across all the layers, consistently outperforming the traditional approach.

5 Towards a glass-based storage system

We have not (yet) built a full storage system and are currently building out (multiple) prototype read and write heads. We follow a clean-state data-driven approach to guide the system design, using large-scale fine-grained event-driven simulations with both synthetic and real-world storage traces. In this section we enumerate the primary design principles of the system and describe some of the most important design choices so far.

There are several factors that heavily influence our design thinking:

Cloud-first Designing solely for the cloud means we are not bound to any form factors or physical constraints that facilitate deployment outside the cloud (such as offices, enterprise data centers, mobile phones, or homes). We use an end-to-end system design to enable us to maximize performance and minimize cost. We co-design the read and write hardware, system hardware and software stack together from the ground up, to create a system that can operate efficiently *only* in the cloud.

Elasticity and Disaggregation Workloads vary over time, often at multiple time-scales. There can be diurnal patterns, and over the longer term the workload can change, for example from being write-dominated to read-dominated. Elasticity means that, over time, only the resources that are needed to service the current system load are used.

The read and write processes differ in three fundamental ways: *(i)* the lasers used for writing are expected to cost an order of magnitude more than the read technology, *(ii)* they use different technologies, and *(iii)* the physical size of the hardware. Disaggregating the read and write heads from the glass media enables us to maximize the utilization rates of the components. It also gives us the opportunity to design a system where we can operate on data in situ and [re-]deploy resources in the system, achieving elasticity, and obviating the need to transfer data between tiers as access characteristics change.

Volumetric storage The glass is a three-dimensional media. The read process concurrently images many voxels in the same XY-plane, and can subsequently image many layers in the Z-dimension. Existing storage technologies (even optical discs with multiple layers) lay out data and read sequentially in a single XY-plane, and do not leverage the third dimension. Because scanning in the Z-dimension for glass storage is fast, we can lay out data in the Z-plane, minimizing XY-seek overhead.

Density In the first system we anticipate that in a volume equivalent to a DVD-disk we can write about 1 TB. The technology can potentially get to 360 TB [21]. The motivation to increase density is not driven by media cost per se, as in most storage technologies, as the cost of the media is negligible. Increasing density improves read performance because of the volume of data in the field of view when reading.

6 Conclusion

We are at beginning of a new era, where we will see zettabytes of data being stored in the cloud. Existing storage technologies have properties that make them less than ideal for cloud storage, and in Microsoft's Project Silica, Microsoft Research and Southampton University, are exploring if quartz glass might be the future media for mass storage in the cloud.

References

[1] Bairavasundaram, L. N., Goodson, G. R., Pasupathy, S., and Schindler, J. An analysis of latent sector errors in disk drives. In *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (New York, NY, USA, 2007), SIGMETRICS '07, ACM, pp. 289–300.

[2] Balakrishnan, S., Black, R., Donnelly, A., England, P., Glass, A., Harper, D., Legtchenko, S., Ogus, A., Peterson, E., and Rowstron, A. Pelican: A building block for exascale cold data storage. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)* (Broomfield, CO, 2014), USENIX Association, pp. 351–365.

[3] Brewer, E., Ying, L., Greenfield, L., Cypher, R., and T'so, T. Disks for data centers. Tech. rep., Google, 2016.

[4] Curtis, K., Dhar, L., Hill, A., Wilson, W., and Ayres, M. *Holographic Data Storage: From Theory to Practical Systems*. Wiley Publishing, 2010.

[5] Digital, W. Western digital unveils next-generation technology to preserve and access the next decade of big data. `https://www.wdc.com/about-wd/newsroom/press-room/2017-10-11-western-digital-unveils-next-generation-technology-to-preserve-and-access-the-next-decade-of-big-data.html`, 2017.

[6] Grawinkel, M., Nagel, L., Mäsker, M., Padua, F., Brinkmann, A., and Sorth, L. Analysis of the ecmwf storage landscape. In *Proceedings of the 13th USENIX Conference on File and Storage Technologies* (Berkeley, CA, USA, 2015), FAST'15, USENIX Association, pp. 15–27.

[7] Hesselink, L., Orlov, S. S., and Bashaw, M. C. Holographic data storage systems. *Proceedings of the IEEE 92*, 8 (Aug 2004), 1231–1280.

[8] Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag. 29*, 6 (2012), 82–97.

[9] IDC. Data age 2025. https://www.seagate.com/our-story/data-age-2025/, 2017.

[10] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2014), CVPR '14, IEEE Computer Society, pp. 1725–1732.

[11] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (USA, 2012), NIPS'12, Curran Associates Inc., pp. 1097–1105.

[12] Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[13] Pawley, J. *Handbook of Biological Confocal Microscopy*. Cognition and Language. Springer, 1995.

[14] Pinheiro, E., Weber, W.-D., and Barroso, L. A. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies* (2007), FAST '07, USENIX Association.

[15] Schroeder, B., and Gibson, G. A. Disk failures in the real world: What does an mttf of 1,000,000 hours mean to you? In *Proceedings of the 5th USENIX Conference on File and Storage Technologies* (Berkeley, CA, USA, 2007), FAST '07, USENIX Association.

[16] Seagate. Hamr technology. `https://www.seagate.com/www-content/ti-dm/tech-insights/en-us/docs/TP707-1-1712US_HAMR.pdf`, 2017.

[17] Shimotsuma, Y., Sakakura, M., Kazansky, P. G., Beresna, M., Qiu, J., Miura, K., and Hirao, K. Ultrafast manipulation of self-assembled form birefringence in glass. *Advanced Materials 22*, 36, 4039–4043.

[18] Shribak, M., and Oldenbourg, R. Techniques for fast and sensitive measurements of two-dimensional birefringence distributions. *Appl. Opt. 42*, 16 (Jun 2003), 3009–3017.

[19] Wood, R., Williams, M., Kavcic, A., and Miles, J. The feasibility of magnetic recording at 10 terabits per square inch on conventional media. *IEEE Transactions on Magnetics 45*, 2 (Feb 2009), 917–923.

[20] Zhang, J., Gecevičius, M., Beresna, M., and Kazansky, P. G. Seemingly unlimited lifetime data storage in nanostructured glass. *Phys. Rev. Lett. 112* (Jan 2014), 033901.

[21] Zhang, J., Čerkauskaitė, A., Drevinskas, R., Patel, A., Beresna, M., and Kazansky, P. G. Eternal 5d data storage by ultrafast laser writing in glass. *Proc.SPIE 9736* (2016), 9736 – 9736 – 16.

[22] Zhu, J. G., and Wang, Y. Microwave assisted magnetic recording utilizing perpendicular spin torque oscillator with switchable perpendicular electrodes. *IEEE Transactions on Magnetics 46*, 3 (March 2010), 751–757.

[23] Zhu, J. G., Zhu, X., and Tang, Y. Microwave assisted magnetic recording. *IEEE Transactions on Magnetics 44*, 1 (Jan 2008), 125–131.