Could cloud storage be disrupted in the next decade?

Andromachi Chatzieleftheriou, Ioan Stefanovici, Dushyanth Narayanan
Benn Thomsen, Antony Rowstron
Microsoft Research

Abstract

If you were asked today *"What are the three dominant persistent storage technologies used in the cloud in 2020?"*, you would probably answer HDD, flash and tape. If you were asked this question in 2010 you would have probably answered HDD, flash and tape. Will this answer change when you are asked this question in 2030?

1 Introduction

The ability to store and retrieve data is fundamental; since the beginning of computing when punch cards and paper tape were used, the technologies used to store data have been critical to the success of computing. In the cloud era, we are seeing an unprecedented demand for storage capacity and for different tiers with different price/performance trade-offs. There has been so much innovation in the last decade in compute and networking, much of it driven by the needs and scale of the cloud. Yet, we have seen little *fundamental* innovation in storage. There are three primary storage technologies: flash, hard disk drives (HDD) and magnetic tape. Is this about to change in the next decade?

All successful technologies follow an innovation S-curve [43] (Figure 1). The x-axis is time and the y-axis is a metric of interest (e.g. GBs/$ or IOPS/$). There are three phases; the *era of ferment* is when the technology is nascent. An early technology is slow to start, and little progress is made on the metrics of interest in this stage. The curve *takes off* when the technology is tamed, with the basic mechanisms (or physics) understood well enough to enable rapid scaling. Typically, the scaling is captured by a rule of thumb such as a doubling of capacity every three years. Finally, in the *maturity* phase it is no longer possible to maintain the rate of improvement using the mechanisms exploited during take-off, and the technology is often said to have *rolled over*.

In Figure 1 there is a red vertical line; when a technology passes this point in its life cycle it is likely that a *discontinuity* will occur, meaning that a new technology

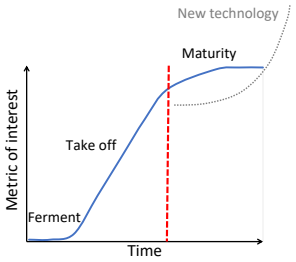


Figure 1: The innovation S-curve.

S-curve will emerge, as shown in the gray line. Initially, the performance of the new technology will be below that of the incumbent technology (when it is in its era of ferment), but *if it is able to reach the take-off phase*, it can rapidly surpass the incumbent. There are many classic examples; for instance for consumer music, cassette tapes killed vinyl records, CDs killed cassette tapes, flash storage in MP3 players killed CDs, and now streaming is killing flash storage. In this paper we use the S-curve as a *qualitative*, rather than quantitative, way of understanding the technology life cycle.

2 The Cloud Disruption

Mature technologies might survive for many decades with only incremental innovation. However the cloud is changing the storage landscape in several ways that make this unlikely in the future:

The curse of exponential growth. Cost is an overriding concern in the cloud. To compete, cloud providers must cut costs, of which hardware is a large part. The size of cloud providers lets them push these cost concerns down to hardware vendors who must "make it up in volume". To sustain exponentially-growing demand without incurring exponentially-growing costs, cloud providers will only buy storage hardware that is also dropping exponentially in cost per gigabyte. Storage technologies that

are in the maturity phase will find this increasingly hard.

Tier virtualization. Storage tiers in the cloud are virtualized; a customer buys a Service-Level Agreement (SLA) rather than a specific storage technology. This lets the provider choose the mix of storage media that can support the SLA at the lowest cost. Since the cost and performance of each storage medium changes as it migrates over its S-curve, the provider can re-optimize this mix transparently to the end-user. For example, as the capacity per dollar of flash increases, it starts to displace HDD storage workloads. Similarly, HDDs are displacing tape workloads. A technology whose growth (in GB per $) slows will have its workload share and market share stolen by other technologies, both current and new.

Increased utilization. The cloud centralizes storage resources. We believe that due to multiplexing many workloads and dynamic provisioning, the average media capacity utilization in the cloud is higher than in other scenarios. Individual device capacities are also higher in cloud storage. It is quite normal to deploy 14TB+ HDDs, to provision capacity dynamically quarter-to-quarter, and to target high utilization rates to reduce costs. Hence, while the total volume of data being stored per year is clearly increasing, the consequence of increased utilization is that the absolute volumes of units shipped may shrink in the short to mid-term. For some media, the market volume in units may become too small to be economically viable to invest significant amounts of money, and the technology will enter the *maturity* phase.

Sustainability In the cloud era, sustainability is of increasing importance. Media lifetime is critical because old media needs to be replaced periodically to ensure the readability of the data. Storage media containing customer data cannot be taken off-site [8] and are destroyed on-site. Also, the tight integration in current devices makes it impossible to service them in the field or to recycle components. As a result, the total cost of ownership increases with the age of the data. Additionally, the data migration during the *refresh cycle* has a significant impact on the system resources (e.g., storage bandwidth).

Legacy form factors At cloud scale, meeting workload demands while maximizing the utilization of all system components is crucial, as cloud providers ultimately have to absorb the cost of underutilized resources. Disaggregation [2, 4, 18, 20, 24, 25, 29] allows the load to be balanced across millions of devices; but it does not reduce the wastage of resources *within the device*. For example, HDDs [9] are fundamentally limited in the IOPS/TB they can provide per-unit. To meet the IOPS requirements, cloud providers are forced to buy more HDDs, often resulting in *stranded capacity*. This is a consequence of the tight integration within the device to fit a legacy form factor such as the 3.5" HDD; it is difficult to customize the ratio of head count to capacity without a huge investment in changing existing highly optimized production lines. By contrast, cloud storage has no intrinsic dependency on legacy form factors. The smallest unit of hardware deployment is the rack and the basic requirement is that a unit be compatible with the loading docks and power budget of the data center. This affords a tremendous amount of design freedom in hardware form factors for new cloud-first storage technologies.

Specialization at scale Cloud storage is currently designed at the exabyte-scale and will soon be designed at the zettabyte-scale. At this scale, in a very cost-conscious operating environment, with tier virtualization and the ability to ignore legacy form factors, deploying novel storage media is very feasible. A large cloud provider has the scale to bootstrap a new storage technology independently, with sufficient demand to make it commodity-priced and to generate the production volume that enables *take off*. To put this into context, a state-of-the-art HDD storage rack provides approximately 10 PB of raw storage, making 1 EB $\sim$100 racks or approximately 60,000 HDDs. This volume is sufficient to drive costs down, and for a cloud that spans hundreds of data centers around the world, this is a very manageable deployment size for a new technology.

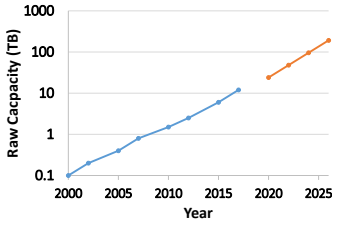
Tail latency Cloud services have very little insight into the workloads that generate the storage load. When provisioning enterprise storage, the type and nature of the workloads that run against it can be considered. In cloud storage a service simply provides an SLA, irrespective of the workload the customer is running. This has forced cloud providers to build systems that control tail latency, as this impacts SLAs [16]. This software-only approach does not solve the problem entirely: we need better hardware support [9]. However, most incumbent storage technologies were designed for throughput, and retrofitting support for tail latency is difficult.

Considering the factors above, incumbent storage technologies could be facing a **perfect storm** of challenges. We believe the time is ripe for one or more of them to be disrupted by new ones. These new technologies are most likely to succeed if they are designed *cloud-first*, meaning the hardware and software stack are designed from the ground up for cloud workloads and for the specific constraints and challenges outlined above.

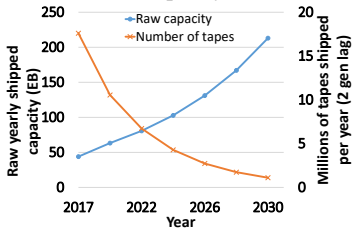
In the next three sections we examine the three incumbent storage technologies, looking at historical trends and why we believe they are hitting their limits. We consider them in the order we believe they are most vulnerable to disruption: magnetic tape, HDDs, and flash.

3 Magnetic tape

Magnetic tapes are widely used for archival storage. Contrary to expectations [23], they are not dead yet, due to exponentially growing demand for cloud storage ca-



(a) LTO **raw** capacity versus time



(b) Capacity and tapes per year.

Figure 2: Magnetic Tapes

pacity. Compared to flash and HDDs, tape is at a different point on its innovation curve. Figure 2a shows in blue the *raw* Linear Tape Open (LTO) [27] capacity increases since 2000, and extrapolates these out to 2025 in orange [10, 37]. The raw tape capacity almost doubles every two years. However, to maintain this trend, tape is reaching the point where technological innovations are needed. Recently IBM announced sputtered tapes, a new type which would allow for higher tape capacities [19].

One big challenge for the tape industry is that it needs to service multiple sectors that are evolving independently in different ways. For example, a significant fraction of users only requires a small scale setup consisting of a limited number of drives and tapes used for archiving. With both cool and nearline cloud storage being relatively cheap and HDD capacities hitting more than 15 TB, the economic benefits of using tapes at this scale are no longer compelling.

The other challenge is that the market for tapes appears to be growing slower than the tape capacity. This means the tape volume shipping each year will potentially decline. In 2017 the industry reported that 109 EB of compressed LTO storage capacity were shipped in total, corresponding to 18 million tapes [38]. However, the total capacity shipped is just 12.9% higher than the previous year. Recall that every two years tape capacities almost double, so on average the tape capacity is increasing by 40% per year. The impact of this is that while the total capacity shipped is rising, the total number of tapes shipped is going down [39]. Figure 2b shows the projected tape volume shipped per year, assuming the 12.9% increase continues each year. By 2030 the market would be approaching 210 EB of raw capacity per year. Figure 2b also shows the number of tape cartridges shipped each year, which would drop to almost 1 million by 2030. In this figure we have optimistically assumed that two generation old tapes dominate the year, otherwise tape shipping could drop even faster. As the number of tapes gets smaller, the drive sales are also dropping. Undoubtedly, the market is shrinking, which explains why there are only two LTO manufacturers today: Fujifilm and Sony.

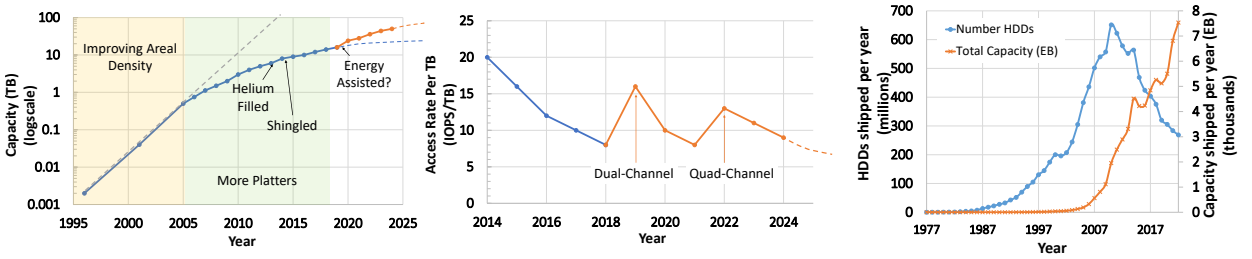
In the cloud context, magnetic tape technology has the most compromises among the three incumbent technologies. To increase drive and media sales, the tape industry has made each drive compatible with two generations of tape media. Despite the media lifetime which is touted as a decade or longer, ensuring the readability of the media can be a challenge as new tape generations are released almost every two years. Anecdotally, several organizations feel the need to migrate their tape storage every 6 or less years simply to ensure they can read their data in the future [11]. This means maintaining many generations of a technology and migrating data to new media, which is expensive. Also, despite the improvements in tape capacity and drive throughput, the basic library design has changed very little. The libraries and tapes themselves are prone to environmental conditions such as humidity, temperature and dust, which complicates their deployment and maintenance, while library robot failures are also rather common.

Summary It is unclear whether magnetic tape is entering its *maturity* phase yet for raw capacity. However, fundamental innovation is needed to maintain current momentum and the economic incentive to innovate decreases as the market is shrinking. The increasing cost of preserving archived data on tapes (which increases with the age of the data), along with exponential growth in the demand, questions the future of tapes over the next decade.

4 Hard Disk Drives

Figure 3a shows the progression of HDD capacity since 1995 [13], where the y-axis is logarithmic. From 1995 to 2005 the hard disk drive industry was primarily exploiting increases in areal density, doubling the capacity of individual drive units around every 2 years.

Around 2005, the industry's ability to continue scaling areal density at this rate began to slow. To compensate, they started increasing the number of platters per drive, with a state-of-the-art disk today having around 9 platters. Having maximized the number of platters in a standard HDD form factor, helium-filled drives (around 2012) allowed platters to be thinner and heads to fly closer to platter surfaces, as helium is less dense than air. Shingling gave a further 20% capacity boost at the cost of removing random write capabilities. For the last decade or so, the industry has been promising to move from today's Perpendicular Magnetic Recording (PMR) write technology to Energy-Assisted Magnetic Record-



(a) HDD capacity roadmap

(b) HDD IOPS/TB roadmap

(c) HDD capacity and units shipped / year

Figure 3: Hard Disk Drives

ing (EAMR), which would allow the rate of areal density scaling to increase again. Multiple options are being explored, including Heat Assisted Magnetic Recording (HAMR) [33] and Microwave Assisted Magnetic Recording (MAMR) [17], with each manufacturer championing a different one. However, operationalizing these technologies has proven both complex and expensive.

Figure 3a shows a classic innovation S-curve, for HDD capacity. People tend to use capacity as a proxy for GB/$, because if this does not fall with each new generation, then customers will not move from their existing drives to the new generation! The figure clearly shows that the technology has reached *maturity* and is beginning to roll over in terms of capacity. GB/$ increases have slowed down, and a significant amount of investment in new HDD technologies is needed to achieve them at the same rates as before. However, this tells only part of the story. The other important metric of interest for HDDs is IOPS/TB.

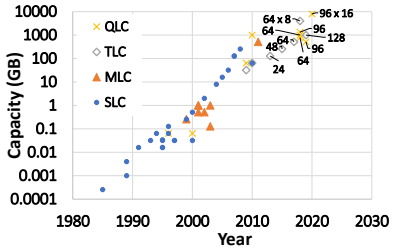
Figure 3b shows IOPS/TB over time since 2014 [6], which have dropped significantly in the last few years. As the industry started relying on adding more platters to increase unit capacity, the average per-head utilization (and by extension throughput) has decreased. This may seem counter-intuitive, but it's important to remember that only a single platter can be accessed at any given time. Current HDD designs time-multiplex a single actuator between the different platters' heads, as servo tracks unique to each platter are followed separately by each platter's head. Changing this design to increase the level of parallelism per-actuator would incur a significant financial cost or increased technical complexity, and would only improve sequential IO performance. This means that even though per-drive capacity has increased, effective access to it has decreased. If the IOPS/TB get too low, it becomes infeasible to use the entire drive capacity for regular online storage. Cloud providers tend to use only part of the disk for online storage, then use the remaining capacity (often referred to as stranded capacity) to provide lower-performance (cool/nearline) storage, at a lower cost to the customer. The balance between these tiers depends on the IOPS/TB provided by the drives. Dual actuators were introduced in 2019 [32], effectively doubling the IOPS/TB. However, as per-unit capacity of HDDs continues to increase, this is a losing battle. Quad actuators may be introduced in a few years, providing a further boost, however going beyond four seems very challenging. Mitigating these problems requires complex targeted solutions (e.g.: multiple actuators, blending of both hot and cold data on the same devices), rather than a trend that brings periodic improvements at steady rate.

Figure 3c shows the number of HDDs shipped per year (in millions) [35], along with the total capacity (in 1,000 EBs) that would have been shipped if every disk was at the maximum capacity available that year. The datapoints for 2020-2022 are predictions. There are several trends causing tension here. The HDD industry needs to increase the per-unit capacity to keep market share, and while the total capacity demand is increasing, it is not doing so at a rate that keeps the number of units sold each year from decreasing. By 2022, the volume of units is predicted to be just half of its peak a few years earlier. This means that any fixed production costs that HDD manufacturers incur have to be split over fewer units, pushing the fixed cost per unit higher.

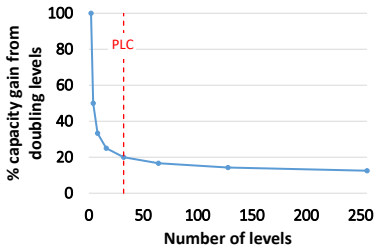
Summary HDD technology has reached *maturity* and its capacity and performance are *rolling over*. The costs and technical complexity of increasing capacity per unit will continue to increase, and technological breakthroughs are needed to keep HDDs affordable in the cloud.

5 Flash

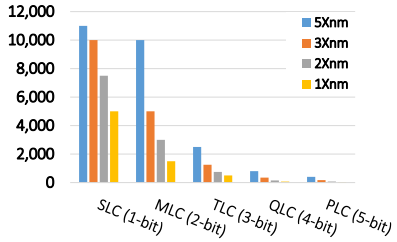
To understand where flash is on its S-curve, let us consider how its areal density has been increasing over time. Areal density (measured in, e.g., Tb/in^2) determines the amount of storage capacity in a given chip area and therefore largely determines the cost per gigabyte of the technology. The obvious way to improve areal density is to decrease *feature size*, i.e., increasing the number of cells in a given chip area by decreasing the cell size. This is attractive as the gains are *quadratic*: a 10% reduction in



(a) Capacity over time [42]



(b) Impact of scaling bits per cell



(c) P/E cycle endurance as a function of cell bit levels and process size [28]

Figure 4: Flash

feature size gives 20% more density, and halving the feature size quadruples the density. Unfortunately this has hit the same limit as other CMOS based technologies and feature sizes are no longer shrinking. Figure 4a shows the historical capacity growth of NAND flash chips.

Until around 2007, manufacturers were able to improve the areal density of SLC (Single Level Cell, 1 bit per cell) flash and did not need other scaling techniques. Since 2007 we have seen significant efforts on increasing the number of bits per cell rather than the number of cells per unit area. MLC (Multi-Level Cells) have 2 bits per cell and date from around 2000. Since 2009 we have seen TLC (Tri-Level Cells, 3 bits per cell) and recently QLC (Quad-Level Cells, 4 bits per cell). PLC (Penta-Level Cells, 5 bits per cell) flash is on the roadmap [36]. Growing capacity in this way is more challenging than increasing areal density, as adding each bit requires doubling the number of levels that a single cell supports. As the number of bits increases, the relative gain of adding another bit drops rapidly as shown in Figure 4b. Going from SLC to MLC doubled the capacity, whereas going from PLC to 6 bits per cell would only increase capacity by 20%. We are thus at the point where investing in more bits per cell results in only modest gains.

Increasing the number of levels also reduces the write endurance of the flash which impacts device lifetime and TCO (total cost of ownership). Figure 4c shows the impact on endurance of both increasing the number of bits per cell and also reducing the process size (figures for PLC are estimates). The typical solution has been to reduce capacity utilization by overprovisioning capacity in the device. This does not directly improve cell endurance but it reduces the number of device-level writes by reducing *write amplification*, the ratio of device level writes to application writes caused by garbage collection. This can have a significant effect initially, e.g., increasing the overprovisioning from 5% to 50% gives a 30% loss in usable capacity but a 5x reduction in write amplification. Further overprovisioning has diminishing returns, e.g. going from 50% to 100% gives a 25% loss in capacity for a less than 20% reduction in write amplification [31]. When overprovisioning exceeds the capacity gain from more bits, it stops making sense. Providers currently manage the write workload to flash tiers to avoid devices failing too soon. At the same time they are trying to increase server lifetimes (e.g., from 3 to 5 years) as the Moore's Law benefits of frequent upgrades have diminished. A decreasing trend in endurance will make this harder.

From 2011, the industry has increasingly relied on *3-D NAND*, also referred to as V-NAND, for density improvements by increasing the number of 3-D layers. This gives a *linear* benefit in density with the number of layers, i.e. not as good as feature size reduction but significantly better than adding bits per cell. Figure 4a shows the 3-D NAND based data points annotated with the number of 3-D layers. Scaling the number of layers further is challenging [41] and will require significant improvements in the 3-D NAND process. One of the challenges is that memory channels must be etched between the 3-D stacks of bits to address them. As the number of layers increases so does the aspect ratio of these channels. This increases the occurrence of defects [34]. The difficulty of scaling 3D NAND is why the industry continues to add bits per cell despite the diminishing capacity gains.

Recently manufacturers have resorted to "die-stacking" several decks of 3-D NAND to achieve greater areal density (shown as "x 8" or "x 16" in the figure). While this avoids the challenges of increasing the number of 3-D layers in a single deck, it is essentially stacking several chips on each other and does not reduce the fundamental cost per bit.

Summary Flash is a technology approaching the *mature* end of its S-curve. It is difficult to see how exponential growth in flash density can be sustained over the next decade. If the cost per gigabyte will not drop significantly, the cost at the cloud scale will also rise exponentially with the exponential growth in demand.

6 Looking to the future

We have so far considered the different dominant media in the cloud and discussed some of the challenges they face. What technologies can we see today that may displace the incumbents?

Tape could be displaced by glass [3] in the mid-term and/or DNA [7] in the longer term. Glass offers very stable long-term media without bit-rot, and one that will not require active storage management, e.g., media scrubbing or environment management (temperature, cooling, humidity etc.). DNA offers media that can potentially have very high storage density, if the correct storage systems can be designed around the media properties. HDDs could be displaced by flash, or perhaps there will be a resurgence of an old technology in the new cloud era. For example, holographic storage [15] has long promised disk-like capacities at a reasonable cost. In the cloud context, with Moore's Law improvements in the underlying technologies (e.g. digital cameras), innovation in optical components and a deeper understanding of garbage collection, could it be done with no mechanical movement and therefore offer higher IOPS and reliability? Persistent memory technologies such as memristors [1] have long promised to displace flash but have not taken off. Recently, 3-D Xpoint [21] has been developed, with the current strategy being to *emulate* flash or DRAM by providing an NVMe SSD or a DDR DIMM. However, it has much lower density than flash, and worse latency and endurance than DRAM, and cannot transparently replace either. Byte-addressable persistence [12, 14, 40], often considered a selling point for persistent memories, can be built at cloud scale from batteries, DRAM, and flash [22, 30]. Of course, it is possible that some technology that we do not yet think of as a storage medium could be the future!

The design of storage systems using new media will be as important as the media. Our experiences are teaching us the importance of clean slate design of storage systems from the media up, and specifically of *complete co-design* of all aspects, including the media, controllers, hardware interfaces, and the software stack. We have also learned that focusing on a single domain (the cloud) removes the compromises that can make a technology unsuccessful. Complete co-design goes hand-in-hand with another key principle that we refer to as *full disaggregation*. Storage in the cloud is often described as being disaggregated, meaning that the compute servers and storage servers are independent. We consider this the first stage of disaggregation and call it *infrastructure disaggregation*. We view rack level or podset level disaggregation for storage [5, 26] as the second stage, where storage servers and drives are disaggregated into dynamically configurable resource pools; we call this *hardware disaggregation*. In truly cloud-first end-to-end storage systems where you can completely co-design the media, hardware and software, we have what we consider to be the third stage: *full disaggregation*. For each and every resource *and functionality* we consider how to design the storage system to enable elasticity in resource usage to maximize utilization but without sacrificing maintainability. Infrastructure and hardware disaggregation emulate existing interfaces, e.g., by presenting remote devices as local ones, while retaining legacy monolithic software stacks running as single processes on servers. These cannot realize the benefits of full disaggregation, and incur the additional capital and operating costs of servers to run the stacks. To get the benefits of full disaggregation we need to redesign the software stack along with the hardware, i.e. complete co-design.

Key to this is *multi-disciplinary teams* that can innovate across traditional boundaries: across materials, devices, hardware and software. Traditionally different research groups work *independently* on each layer. A physics department works on new media (e.g. resistive memories), an EE department on packaging (e.g. memory controllers), and a CS department on new software stacks (e.g. persistence abstractions). This isolation leads to an *emulation* approach as in the 3-D Xpoint example, rather than true co-design.

7 Conclusion

We started by asking if today's dominant cloud storage technologies would be the same in a decade. They may be, but if there is to be change, then the storage research community should be at its forefront. Much of the community has focused on taking the basic storage hardware *as a given* and optimizing the software stack for it. This *software systems research* is necessary to help the existing technology reach full maturity but insufficient to create a new technology S-curve. We will need fundamentally new technologies from the media up. Systems-level solutions, e.g. RDMA or disaggregation, improve efficiency in the worst case by a few percent and by a small factor at best. *By themselves* they cannot provide the sustainable exponential growth that cloud storage relies on.

The traditional areas of scheduling, garbage collection, data placement, and fault tolerance will take on new challenges and dimensions with new hardware. There are also new issues: what are the right software/hardware interfaces once legacy interfaces are discarded? Which software processing needs to be physically co-located with media and access hardware, and what can be disaggregated? What mix of general-purpose and specialized compute should we build into the design? We invite the storage research community to join us in thinking broadly on how to lead the disruption in cloud storage.

References

[1] The memristor revisited. *Nature Electronics*, 1(5):261–261, May 2018. doi:10.1038/s41928-018-0083-3.

[2] Ganesh Ananthanarayanan, Ali Ghodsi, Scott Shenker, and Ion Stoica. Disk-locality in datacenter computing considered irrelevant. In *Proceedings of the 13th USENIX Conference on Hot Topics in Operating Systems*, HotOS'13, page 12, USA, 2011. USENIX Association.

[3] Patrick Anderson, Richard Black, Ausra Cerkauskaite, Andromachi Chatzieleftheriou, James Clegg, Chris Dainty, Raluca Diaconu, Rokas Drevinskas, Austin Donnelly, Alexander L. Gaunt, Andreas Georgiou, Ariel Gomez Diaz, Peter G. Kazansky, David Lara, Sergey Legtchenko, Sebastian Nowozin, Aaron Ogus, Douglas Phillips, Antony Rowstron, Masaaki Sakakura, Ioan Stefanovici, Benn Thomsen, Lei Wang, Hugh Williams, and Mengyang Yang. Glass: A new media for a new era? In *10th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 18)*, Boston, MA, July 2018. USENIX Association. URL: https://www.usenix.org/conference/hotstorage18/presentation/anderson.

[4] Microsoft Azure Storage. URL: https://azure.microsoft.com/services/storage.

[5] Shobana Balakrishnan, Richard Black, Austin Donnelly, Paul England, Adam Glass, Dave Harper, Sergey Legtchenko, Aaron Ogus, Eric Peterson, and Antony Rowstron. Pelican: A building block for exascale cold data storage. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 351–365, Broomfield, CO, October 2014. USENIX Association. URL: https://www.usenix.org/conference/osdi14/technical-sessions/presentation/balakrishnan.

[6] James Borden and Timothy Walker. HDD Parallelism for Lower TCO: Dual Actuator Implementation. In *Storage Developer Conference*, September 2018. URL: https://tinyurl.com/y9d37uke.

[7] James Bornholt, Randolph Lopez, Douglas M. Carmean, Luis Ceze, Georg Seelig, and Karin Strauss. A DNA-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '16, page 637–649, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2872362.2872397.

[8] Pishoy Bous. Data security Q&A with John Molesky, Azure Security Engineering. URL: https://tinyurl.com/tmyqfna.

[9] Eric Brewer, Lawrence Ying, Lawrence Greenfield, Robert Cypher, and Theodore T'so. Disks for data centers. Technical report, Google, 2016. URL: https://research.google/pubs/pub44830.

[10] 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.

[11] Germ Cancio, Vladim Bahyl, Daniele Francesco Kruse, Julien Leduc, Eric Cano, and Steven Murray. Experiences and challenges running CERN's high capacity tape archive. *Journal of Physics: Conference Series.*, (4), 2015.

[12] Joel Coburn, Adrian M. Caulfield, Ameen Akel, Laura M. Grupp, Rajesh K. Gupta, Ranjit Jhala, and Steven Swanson. NV-Heaps: Making persistent objects fast and safe with next-generation, non-volatile memories. In *Proceedings of the 16th International Conference on Architectural Support for Programming Languages and Operating Systems, (16th ASPLOS'11)*, pages 105–118, Newport Beach, CA, USA, March 2011. ACM Press. doi:10.1145/1961295.1950380.

[13] Wikimedia Commons. Hard Drive Capacity Over Time. URL: https://tinyurl.com/sllt3l3.

[14] Jeremy Condit, Edmund B. Nightingale, Christopher Frost, Engin Ipek, Benjamin C. Lee, Doug Burger, and Derrick Coetzee. Better I/O through byte-addressable, persistent memory. In Jeanna Neefe Matthews and Thomas E. Anderson, editors, *Proceedings of the 22nd ACM Symposium on Operating Systems Principles 2009, SOSP 2009, Big Sky, Montana, USA, October 11-14, 2009*, pages 133–146. ACM, 2009.

[15] Hans J Coufal, Demetri Psaltis, and Glenn T Sincerbox, editors. *Holographic Data Storage*, volume 76 of *Springer Series in Optical Sciences*. Springer, 2000. doi:https://doi.org/10.1007/978-3-540-47864-5.

[16] Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56:74–80, 2013. URL: https://tinyurl.com/y8agod44.

[17] Western Digital. The need for energy-assisted recording technology. In *Western Digital Blog*, October 2017. URL: https://tinyurl.com/qq3mz7o.

[18] Aleksandar Dragojevic, Dushyanth Narayanan, Miguel Castro, and Orion Hodson. FaRM: Fast remote memory. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2014)*, April 2014. URL: https://www.microsoft.com/en-us/research/publication/farm-fast-remote-memory/.

[19] Simeon Furrer, Mark A. Lantz, Peter Reininger, Angeliki Pantazi, Hugo E. Rothuizen, Roy D. Cideciyan, Giovanni Cherubini, Walter Haeberle, Evangelos Eleftheriou, Junichi Tachibana, Noboru Sekiguchi, Takashi Aizawa, Tetsuo Endo, Tomoe Ozaki, Teruo Sai, Ryoichi Hiratsuka, Satoshi Mitamura, and Atsushi Yamaguchi. 201 Gb/in^2 recording areal density on sputtered magnetic tape. *IEEE Transactions on Magnetics*, 54(2):1–8, February 2018. URL: http://ieeexplore.ieee.org/document/7984852/.

[20] Peter X. Gao, Akshay Narayan, Sagar Karandikar, Joao Carreira, Sangjin Han, Rachit Agarwal, Sylvia Ratnasamy, and Scott Shenker. Network requirements for resource disaggregation. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 249–264, Savannah, GA, November 2016. USENIX Association. URL: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/gao.

[21] Jim Handy. Understanding the Intel/Micron 3D Xpoint Memory. In *Storage Developer Conference*, September 2015. URL: https://tinyurl.com/yctjvnwe.

[22] Shaun Harris. Microsoft reinvents datacenter power backup with new Open Compute project specification, March 2015. URL: https://tinyurl.com/yam93hpa.

[23] Jim Gray. Tape is Dead. Disk is Tape. Flash is Disk. RAM Locality is King. In *CIDR 2007 Gong Show Presentations*, Asilomar, CA, USA, December 2007. URL: http://cidrdb.org/cidr2007/gongshow/1Gray.ppt.

[24] Ana Klimovic, Christos Kozyrakis, Eno Thereska, Binu John, and Sanjeev Kumar. Flash storage disaggregation. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys '16, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2901318.2901337.

[25] Ana Klimovic, Heiner Litz, and Christos Kozyrakis. Reflex: Remote flash ≈ local flash. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '17, page 345–359, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3037697.3037732.

[26] Sergey Legtchenko, Hugh Williams, Kaveh Razavi, Austin Donnelly, Richard Black, Andrew Douglas, Nathanael Cheriere, Daniel Fryer, Kai Mast, Angela Demke Brown, Ana Klimovic, Andy Slowey, and Antony Rowstron. Understanding rack-scale disaggregated storage. In *9th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 17)*, Santa Clara, CA, July 2017. USENIX Association. URL: https://www.usenix.org/conference/hotstorage17/program/presentation/legtchenko.

[27] LTO Ultrium Technology. URL: https://www.lto.org/.

[28] Chris Mellor. WD: Storage class memory will not replace DRAM or NAND. URL: https://tinyurl.com/u5flx6y.

[29] James Mickens, Edmund B. Nightingale, Jeremy Elson, Krishna Nareddy, Darren Gehring, Bin Fan, Asim Kadav, Vijay Chidambaram, and Osama Khan. Blizzard: Fast, cloud-scale block storage for cloud-oblivious applications. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, page 257–273, USA, 2014. USENIX Association.

[30] Dushyanth Narayanan and Orion Hodson. Whole-system persistence. In *Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XVII, page 401–410, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2150976.2151018.

[31] Bill Radke. Overprovisioning in all-flash arrays. In *SNIA Education Tutorial Series*, 2013. URL: https://tinyurl.com/tnedryj.

[32] Seagate. Highest Performance for Highest Rack Space Efficiency. URL: https://tinyurl.com/y94k3u9l.

[33] Seagate. HAMR Technology. In *Seagate Technology Paper*, December 2017. URL: https://tinyurl.com/wmvuqtk.

[34] Harmeet Singh. Overcoming challenges in 3D NAND volume manufacturing. *Solid State Technology*, 60(5):18–21, 2017. URL: https://tinyurl.com/t3rz89x.

[35] Statista. Number of HDDs Shipped Worldwide From 1976 to 2018. URL: https://tinyurl.com/txkv9ar.

[36] TechTarget. Intel roadmap includes faster optane ssds, 144-layer nand. URL: https://tinyurl.com/sljrkya.

[37] LTO Ultrium. LTO Ultrium roadmap through to generation 12. URL: https://tinyurl.com/v5umlvz.

[38] LTO Ultrium. Record breaking amount in total tape capacity shipments announced by the LTO program (2017)., March 2018. URL: https://tinyurl.com/y88au7cc.

[39] LTO Ultrium. The LTO Program: Media Shipment Report for Calendar Year 2017, March 2018. URL: https://tinyurl.com/qwy2ule.

[40] Haris Volos, Andres Jaan Tack, and Michael M. Swift. Mnemosyne: Lightweight persistent memory. In *Proceedings of the 16th International Conference on Architectural Support for Programming Languages and Operating Systems, (16th ASPLOS'11)*, pages 91–104, Newport Beach, CA, USA, March 2011. ACM Press.

[41] Steve Shih-Wei Wang. 3D NAND: Challenges beyond 96-Layer Memory Arrays. In *Coventor Blog*, October 2018. URL: https://tinyurl.com/rejxx4u.

[42] Wikipedia. Flash memory. URL: https://tinyurl.com/d9cgbwj.

[43] Wikipedia. Innovation. URL: https://en.wikipedia.org/wiki/Innovation.