A Call for Research on Storage Emissions

Sara McAllister* Fiodar Kazhamiaka† Daniel S. Berger†
Rodrigo Fonseca† Kali Frost† Aaron Ogus† Maneesh Sah† Ricardo Bianchini†
George Amvrosiadis* Nathan Beckmann* Gregory R. Ganger*
*Carnegie Mellon University †Microsoft Azure

Abstract

Major cloud providers have committed to lowering carbon emissions by 2030 across their datacenters, and research has contributed many ideas on how this may be achieved. However, a major contributor to datacenter emissions has not received enough attention: storage. Storage — everything from file storage to inter-application messaging in datacenters — causes 33% of operational emissions and 61% of embodied emissions in Azure's general-purpose cloud, based on a recent study.

This paper identifies key sources of both operational and embodied emissions within distributed storage in datacenters. We also discuss strategies to reduce storage emissions and their challenges due to storage's fundamentally stateful nature.

1 It's time to talk about storage emissions

The systems research community has been working to reduce datacenter carbon emissions. Existing work focuses primarily on reducing emissions of *general-purpose compute* [25, 39, 70, 71, 123, 130, 131, 134], neglecting a large source of emissions: storage.

While storage has received some attention [70, 86, 87, 99, 128, 143], researchers and practitioners have frequently considered it a less important source of emissions. This could not be further from the truth. Storage comprises a sizable portion of both operational (Scope 2) and embodied (Scope 3) carbon emissions in hyperscale datacenters.[1] Recent data from Azure suggests that storage-related emissions — including storage racks and local storage devices — make up 33% of operational and 61% of embodied emissions. Storage racks alone account for 24% of operational and 45% of embodied emissions [131].

In fact, we identify distributed storage as a dominating contributor to emissions in future datacenters, even given aggressive predictions for the datacenters' AI expansion. It is widely believed that GPUs will emit the most operational carbon (which will be partly powered by renewable energy), but their embodied carbon is not nearly as dominant. For example, we observe that a Nvidia A100 GPU has about the same embodied emissions as a 1.6-17 TB SSD [2] or 2 CPUs [70, 78].

As datacenters continue to target compute emissions and deploy renewable energy, storage will dominate overall datacenter emissions due to storage's embodied emissions. Recent research has heavily optimized compute emissions, but these approaches do not generally apply to storage. Storage has fundamentally different constraints, such as ensuring data durability and availability. Thus, while the high-level techniques — including reducing power consumption, shifting power consumption to regions and times where renewable energy is available [7, 25, 35, 113, 123, 135], using fewer devices [44, 50, 51, 64, 121], and extending device lifetime [97, 130, 132–134] — still apply to storage, they face different challenges and tradeoffs.

For example, extending device lifetime leads to higher device failure rates. Whereas compute can usually just be migrated to a new server, storage is fundamentally stateful. Higher failure rates increase the likelihood of data loss, requiring more capacity for erasure-coding, reducing the benefit of extending device lifetime.

This paper informs the conversation on storage emissions by:

- *Identifying distributed storage racks as a first-order emissions problem:* We identify storage as a necessary target for emissions reductions and break down both operational and embodied emissions in Azure's storage, showing the impact of both SSD and HDD storage servers.
- *Bridging the storage–sustainability knowledge gap:* To help researchers tackle storage emissions, we highlight the characteristics of storage that are relevant for sustainability research geared towards reducing carbon emissions.
- *Highlight opportunities and challenges to reducing storage emissions:* We show the opportunities to reduce emissions — including denser devices, longer lifetimes, and new storage technology — but also identify challenges in deploying these solutions in the datacenter.

2 Birds-eye view of cloud storage

Cloud storage is predominantly backed by distributed storage systems. Most data is permanently stored in *storage servers* grouped into storage racks, separated from compute. While data may be stored on 'local' devices attached to a compute server, this is typically used as a cache. In fact, many VM types do not offer locally-attached SSDs [2, 5, 10]. Thus, our focus is distributed storage.

Cloud data storage today has two media options: *hard-disk drives* (HDD), for storing large amounts of data, and *solid-state drives*, for low-latency data access. SSDs are about 2-4x more expensive per bit than HDDs [1, 4]. For carbon, the difference is even more pronounced — SSDs require 3-10x more embodied emissions per bit [70, 128] and more power per bit (Sec. 3.1).

This section discusses storage from a user's perspective (Sec. 2.1) and the storage server configurations that enable it (Sec. 2.2).

2.1 Storage in the cloud

Cloud users rely on distributed storage to guarantee scalabilty, durability, and high availability. As we consider options to reduce storage emissions, we need to carefully consider their impact on these guarantees.

[1] We adopt the greenhouse gas protocol's definition of emission scopes similar to prior work [70, 71]. Scope 1 is negligible [131], we thus focus on Scopes 2 and 3.
[2] Public estimates for embodied emissions of SSDs have a large variance; we believe is due to a combination of variance across suppliers, different technologies and technology specifications, and imprecise modeling efforts.

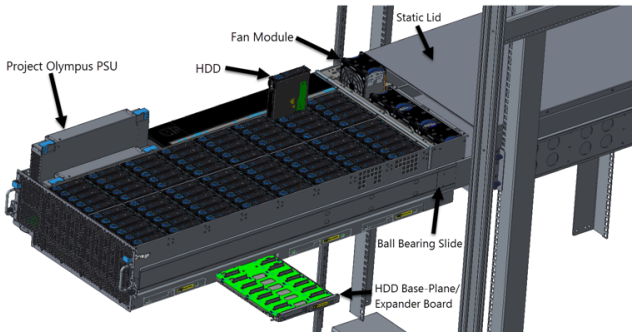


Fig. 1: Project Olympus JBOD ("Just a bunch of disks") blade [74]. Each JBOD blade contains up to 88 HDDs and can store 1.2 PB.

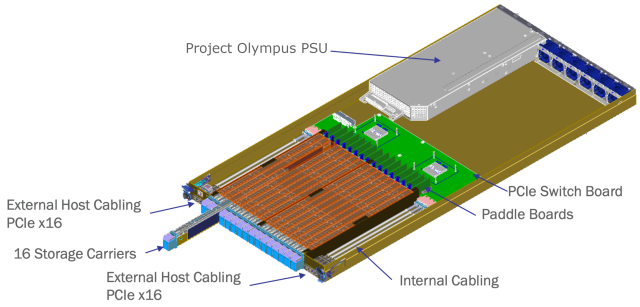


Fig. 2: Project Olympus flash expansion blade [119].

	Size	Blades/Rack	Count	Capacity ('17)	Capacity ('24)
SSD [119]	1U	8	16	128 TB	246 TB [22]
HDD [74]	4U	36	88	1.2 PB	2.6 PB [31]

Table 1: Comparison of Project Olympus' storage servers including the blade size, blades per rack, storage device count, and the capacity of the entire blade from both the original Project Olympus 2017 specifications and updating the storage devices to 2024 capacities.

Scalability. Cloud providers ensure there is enough storage capacity so that cloud applications can scale to store their portion of the zettabytes of data generated annually [118]. To provide this guarantee, cloud providers need to pre-provision storage for the expected data growth, ensuring there is enough extra capacity for additional data.

Durability and high availability. The maxim of storage is that a storage system should not lose data, i.e., the data should be *durable*. Unfortunately, servers fail. At scale, server failures are common [49]. An advantage of distributed storage is data redundancy through replication or erasure-coding [32, 58, 67, 76, 106, 115, 142]. When a storage server fails, the data is recoverable as other storage servers are able to reconstruct the data. Redundant data provides both durability and *high availability* because the data is still generally accessible (i.e., available) through failure. Durability and availability requirements constrain the options to reduce emissions (Sec. 4).

2.2 Storage under the hood

To provide storage products, datacenters have large distributed storage systems [40, 65, 66, 76, 105] that need to be scalable, durable, and highly available. Due to their scale, these storage systems manage large clusters of servers, with hundreds of thousands of storage devices [49, 105]. In this subsection, we dive into the makeup of these storage servers since they are the backbone of distributed storage and the primary source of storage's emissions.

We broadly divide storage servers into two categories — HDD servers (Fig. 1) and SSD servers (Fig. 2) — based on the type of the storage device that provides the bulk of the capacity. As seen in Table 1, HDD servers have more capacity due to their large number of high-capacity devices, but also take up more rack space. SSD servers take up less rack space, but have lower capacities both due to having fewer drives and lower-capacity drives. To reduce emissions in both types of these servers without impacting performance, we need to understand more about the underlying device characteristics.

Hard-disk drives (HDDs). An HDD server's purpose is to store lots of data cheaply. To accomplish this, each server holds many disks (e.g., 88 in Project Olympus, Table 1), referred to as "Just a bunch of disks" or JBODs. These servers store an order-of-magnitude more data per server than SSD servers and about 2.6x more data per rack space. This higher density also helps with emissions (Sec. 4.2).

HDDs contain multiple circular platters that store data magnetically. To write or read data to a platter, the platter's head has to seek to the correct track and wait for the disk to spin to the correct sector. Thus, a key factor in request latency is the speed that the HDD is spinning, i.e., its rotations-per-minute (RPM). RPM affects both wait time and device bandwidth, since the HDD can only transfer data that passes under its active head. Unfortunately, even after significant optimization effort, RPM has not improved much for the past decade [1, 105].

HDDs are growing denser, maintaining their capacity-cost advantage over SSDs. HDDs have grown from one to 20 terabytes over the last decade without changing their form factor [85] through density improvements such as shingled-magnetic recording (SMR) drives that overlap the write tracks to increase bits stored per disk area [27, 137]. The next frontier of HDD density is heat-assisted magnetic recording (HAMR), which allows denser packing of bits by heating disks during writes [24]. HAMR promises to increase device capacities to 50 TB and beyond [120].

Solid State Drives (SSD). SSD servers are smaller and lower capacity than HDDs, but are more performant both due to their lower latency and higher bandwidth [105]. They are offered to cloud users as high-performance storage options [23].

SSDs are built from NAND flash memory. Flash memory has two main limitations: it wears out with writes and does not allow small-granularity overwrites. SSDs have *limited write endurance* — after too many writes, their flash cells can no longer store data [73]. If applications write too much, flash's lifetime can be extremely short, leading to higher embodied emissions. In addition, SSDs have to accommodate flash's lack of overwrites. Flash media in SSDs is organized into large blocks, often gigabytes in size [37, 100]. To overwrite data in a block, the SSD first has to erase the *entire block*, copying any live data in that block elsewhere (a.k.a., garbage collection). These extra writes exacerbate limited flash's write endurance [38, 55, 73, 89, 91, 93, 127].

Like HDDs, SSDs have been getting denser. SSD density has increased through two main mechanisms: increasing the number of layers and increasing cell density. SSDs have been 3D-stacking

Operational Emissions	CPU	DRAM	SSD	HDD	Other
Compute Rack	42%	18%	19%	0%	21%
SSD Rack	32%	8%	**38%**	1%	21%
HDD Rack	26%	5%	7%	**41%**	21%

Table 2: Operational emission breakdown for Azure rack types.

Embodied Emissions	CPU	DRAM	SSD	HDD	Other
Compute Rack	4%	40%	30%	0%	26%
SSD Rack	1%	9%	**80%**	1%	9%
HDD Rack	2%	11%	14%	**41%**	33%

Table 3: Embodied emission breakdown for Azure racks.

layers of cells, growing flash storage "vertically." Today, flash devices can have over 200 layers, and the number of layers is quickly increasing [9, 13]. 3D stacking increases device density but also increases embodied emissions. Flash is also becoming denser by packing bits into cells. Most datacenter SSDs today use tri-level cells (TLC), which store 3 bits per cell. Flash SSDs will soon use quad-level cells (QLC) (4 bits/cell) and penta-level cells (PLC) (5 bits/cell) [110]. Unfortunately, increasing cell density causes lower write endurance, causing quickly diminishing returns. Both of these methods to increase density can improve emissions if they can be successfully deployed (Sec. 4.2).

3 Where do storage emissions come from?

Distributed storage is a large emitter [131]. Unfortunately, there is no standard break down of storage emission sources, which is necessary to understand and reduce storage emissions. Therefore, in this section, we show how each component of storage servers in a datacenter rack contributes to emissions at Azure.

We divide emissions into three components — direct emissions (Scope 1), operational emissions (Scope 2), and embodied emissions (Scope 3) — based on the Greenhouse Gas Protocol's definitions [70, 71, 131]. We do not present Scope 1 emissions since they are negligible [131]. Sec. 3.1 discusses operational emissions, e.g., from power generation, and Sec. 3.2 discusses embodied emissions, e.g., from semiconductor fabs.

We present the emissions from both a SSD storage rack and an HDD storage rack, focusing on the key components (CPU, DRAM, SSD, and HDD). We use the "Other" category to group rack overheads, such as fans, network switches, power supplies, and power delivery units. For embodied carbon, the "Other" category also includes passive material like sheet metal and plastics.

3.1 Operational emissions

Table 2 shows the relative operational emissions of each Azure rack type. To determine energy consumption and therefore operational emissions of different components, we take component energy draws measured under a representative load. Notably, an SSD storage rack has approximately 4× the operational emissions per TB of an HDD storage rack.

Storage devices (SSDs and HDDs) are the largest single contributor of operational emissions. For SSD racks, storage devices account for 39% of emissions, whereas for HDD racks they account for 48% of emissions. These numbers contradict the conventional wisdom that processing units dominate energy consumption [70, 131]: storage servers carry so many storage devices that they become the dominant energy consumers. Thus, the best way to reduce operational emissions in a storage server is to reduce the storage *devices*' energy (Sec. 4).

Since CPUs still cause the next largest portion of the emissions, improving the energy efficiency of CPUs in storage servers may still provide benefits. However, one has to be careful with energy efficiency improvements that increase embodied carbon emissions [128, 131]. For example, advanced semiconductor fabrication nodes reduce operational emissions but increase manufacturing emissions and electricity use [34, 54, 125]. This consideration is particularly important in storage, which is already embodied emission heavy.

3.2 Embodied emissions

We show the relative embodied emissions of each Azure rack type in Table 3. To estimate embodied emissions, we use raw material numbers from vendors, the device's silicon area, and leverage IMEC [8] and Makersite [11] to determine average emissions for manufacturing processes. We ensure that manufacturing and shipping emissions are only counted once and are amortized across components, so that our embodied emissions results are comparable to our operational emissions results.

SSD racks emit approximately 10× the embodied emissions per TB as that of HDD storage racks. The storage devices themselves dominate embodied emissions, accounting for 81% and 55% of emissions in SSD and HDD racks, respectively. While DRAM is the largest embodied emissions contributor in compute servers, this is not true for storage servers due to the many storage devices in these servers. Across both operational and embodied emissions in distributed storage clusters, there is a clear need to reduce emissions from the storage devices themselves.

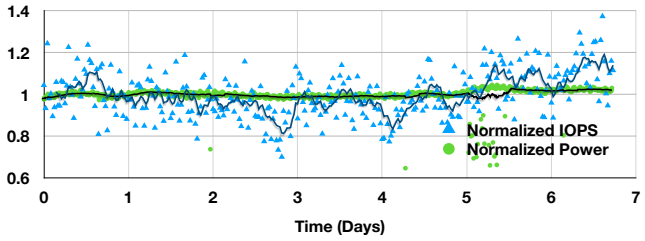
4 Emission reduction in storage

Solutions that effectively reduce carbon in compute servers are not generally effective in storage servers. We now consider the important opportunities and challenges in reducing storage emissions. To structure the discussion, we model emissions by breaking apart operational and amortized embodied emissions:

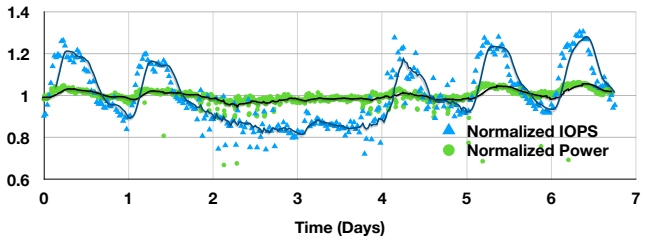
$$\begin{aligned}&\text{Annual Carbon Emissions}\\&= \text{Operational Emissions} + \text{Embodied Emissions}\\&= \sum_{\text{Devices}} \left(\frac{\text{Watt-Hours}}{\text{Device}} \times \frac{\text{Carbon}}{\text{Watt-Hour}} + \frac{\text{Carbon}}{\text{Device}} \times \frac{1}{\text{Lifetime}} \right)\end{aligned}$$

This simple model tells us that emissions can be reduced in five ways: using fewer devices, lowering power, reducing carbon intensity of power, reducing per-device embodied emissions, or increasing server and device lifetime.

Prior work uses three main methods to achieve these reductions in compute servers: (1) reducing and shifting power by increasing utilization [45, 60–62, 79, 94, 95, 101, 111, 124, 141] and moving computation to times/locations with more renewable energy [7, 25, 35, 113, 123, 135]; (2) using fewer, more efficient compute devices and reducing emissions per device [44, 50, 51, 64, 121]; and (3) increasing lifetimes by identifying places where older device performance is adequate [97, 130, 132–134].



(a) HDD cluster



(b) SSD cluster

Fig. 3: Power and IOPs over 7 days. Both power and IOPs are normalized to their average value over 30 days.

In storage, these reductions are harder to accomplish because: (1) lowering and shifting power consumption relies on power varying with usage, which is not true in storage (Sec. 4.1); (2) reducing devices implies denser storage devices, which have performance and endurance limitations (Sec. 4.2); (3) increasing lifetime increases the failure rates, threatening durability (Sec. 4.3). There are potential solutions to address or mitigate all of these challenges, but we need further research to enable storage emission reductions.

4.1 Lowering and shifting power

Operational emissions can be reduced by shifting work to reduce peak power consumption and aligning power use with the availability of renewable energy. Unfortunately, it is difficult to shift power consumption in storage, and, more importantly, operational emissions are a small fraction of overall emissions in storage. Also, since storage is a relatively small fraction of operational emissions, any operational emissions gains need to not come at the expense of embodied emissions.

Storage power usage is basically flat. Power usage does not vary much in storage clusters, despite high variation in IOPS. To quantify this effect, we plot the power usage of both an HDD cluster and SSD cluster at Azure along with the IOPS in each cluster (Fig. 3). These measurements were collected over 7 days. Each point for power is the average power consumption over a 5-minute window, normalized to the average power. Each point for IOPS is the average IOPS in a 30-minute window, normalized to the average IOPS. We also plot a 4-hour moving average for both IOPS and power.

For HDD clusters, we see that IOPS varies, though not in a diurnal pattern, but power does not. Between late in day 2 to the start of day 3, the IOPS increases by 61%, but the power remains constant. Over the 7 days, power varies little, having a standard deviation of 3%. The lack of power variation is due to HDDs using most of their power to continue spinning, which needs to occur regardless of IO accesses [68, 102, 103].

Power also varies little in SSD clusters. Over 7 days, we observe a standard deviation of 3%. However, SSDs show a more consistent diurnal pattern during weekdays (the trace starts on a Thursday) that we can see mirrored slightly in power. For instance the IO drop on the zeroth day of 28%, results in only a power drop of 0.04%.

Due to the relatively consistent power usage of both SSD and HDD clusters, shifting IO to reduce peaks or to when/where there is more renewable energy will not significantly reduce operational emissions. Conversely, *increasing IO has little impact on operational emissions.*

Prior work aims to reduce hard drives' energy, but does not consider embodied emissions. The observation that HDDs require consistent power is not new; prior work has tried to reduce HDD energy since the early 2000s. This work fits into two categories: *(i)* caching to create idle IO periods and *(ii)* distributing data to enable power proportionality.

HDDs use far less power when in idle or sleep modes [16]. However, HDDs have little IO variation currently in Azure's datacenters, so idle periods would need to be intentionally created. Prior work has different strategies, such as using a relatively low-power disk to store most hot data [41], separating hot and cold data [77, 87, 92, 107], or employing better prefetching and caching policies [112, 122]. These policies often do not work for common maintenance tasks that stream through large amounts of data sequentially, such as scrubbing [33, 117], and are hard to change in cloud environments where much of the work is *not under the cloud provider's control.* As shown in Pelican [33], datacenter power provisioning would also have to change to accommodate variable power, since typically power is provisioned for storage server's peak — when all disks are running.

Alternatively, some work has suggested leveraging data replication to turn off storage servers when either IO is low enough [28, 33, 46, 129] or there is not enough available green energy [86]. Turning off the entire server allows for energy savings from all components, not just disks [63]. Unfortunately, most prior work assumes data *replication*, where the storage system stores multiple copies of data, whereas today's datacenters use more space-efficient erasure codes [81, 84]. One prior paper did consider leveraging redundancy in erasure-code-based distributed storage systems [108], but much more work can be done especially factoring in embodied emissions and the constraints of modern datacenters.

Opportunities to reduce operational emissions. Although storage operational emissions is a smaller part of datacenter emissions, there is still potential to reduce it. SSDs are more power proportional than HDDs [22], meaning some prior HDD work may be more impactful in SSDs. For HDDs, running at lower rotational speeds instead of fully off could reduce power without affecting durability [41, 72], since HDDs lifetime decreases with frequent power cycles [68]. Lower rotational speeds also allow an increase in areal density [69] — synergizing well to also decrease embodied emissions-per-bit, at the expense of latency and bandwidth. The downside is a drop in IO, some of which may be recoverable from deploy dual-actuator HDDs [19, 42, 88, 106].

4.2 Fewer, denser drives

The most direct way to reduce storage emissions is to use fewer, denser storage devices. Denser storage devices could lower power consumption as well as the number of servers and racks required to store the same data, reducing both operational and embodied emissions. This subsection discusses three main ways to increase density: moving from SSDs to HDDs, using denser SSDs and HDDs, and deploying new storage media.

Unfortunately, increasing storage density is not straightforward. Denser devices typically do not have proportionally higher IO, reducing the IO per bit and introducing new performance constraints. Overcoming these constraints requires more research, as does holistically and accurately estimating emission gains from denser storage.

Move from SSDs to HDDs. As shown in Sec. 3, SSD servers emit more than HDD servers, both per-rack and per-bit. To reduce emissions, all applications that can tolerate HDD performance should be moved to HDDs. Unfortunately, clouds today are sometimes locked into cloud users' choices between the two tiers of storage. For instance, Azure and AWS VM-attached disks default to SSD storage [3, 23]. This default increases emissions if cloud applications do not need the additional performance, but it is hard to avoid because the customer has paid to get SSD performance. Moving to HDDs without changing this user-performance agreement would essentially force the provider to still keep the entire working set in a more performant media such as DRAM or SSDs, increasing emissions.

Alternatively, cloud providers could encourage users to choose lower performance, more sustainable storage. One way to encourage movement to more sustainable storage is to increase awareness of the difference in emissions between storage options — whether through pricing or explicit calculations of emissions when allocating storage space. Performing these calculations accurately per workload requires additional research.

Deploying denser SSDs and HDDs. HDD and SSD density is already increasing. Generally, denser drives reduce embodied emissions per bit because they require about the same materials while storing more data [99, 143]. However, generational advances sometimes require more resources and manufacturing emissions.

For HDDs, the next density jump comes from HAMR (Sec. 2.2). HAMR adds both operational and embodied emissions due to the addition of lasers on each write head. These drives also have to use different materials to increase magnetic stability of the bits at room temperature [24], changing the embodied emissions of the hard drive platters, though we expect the density increases to still lower overall embodied emissions. For SSDs, adding layers adds to the manufacturing emissions, increases manufacturing complexity, and lowers reliability [140].

These considerations will add embodied emissions at the device level, but still likely result in a better embodied emissions-per-bit due to the additional density. In order to understand the emissions benefits with denser storage, we need to study the emission impact of the denser technology.

IO prevents full adoption of denser HDDs and SSDs. The additional carbon emissions from denser HDDs and SSDs are not the only challenges for deploying these drives to reduce emissions—denser drives also can perform less IO per TB of capacity.

IO bottlenecks are already becoming a challenge in datacenters for HDDs, primarily because higher-capacity HDDs do not increase their bandwidth. For instance, Seagate has LCA analysis for its Exos HDDs show that its 18 TB HDD has 59.6% fewer kg CO_2e per TB-year compared to its 10 TB drive [20, 21]. However, the 18 TB HDD's bandwidth only increases 8.4% and has no increase in random 4KB IOPS [17, 18]. In order to use the 18 TB drives instead of 10 TB drives, we would need to reduce IO per GB stored. But there is little headroom available — many storage applications already saturate today's HDD bandwidth.

Additionally, both SSDs and HAMR HDDs suffer write-endurance problems. SSDs' limited write endurance is a well-known problem [26, 30, 143]. Write endurance gets worse with higher cell density. PLC is projected to have 16% of the write endurance of today's TLC drives [15]. HAMR drives add a laser on each read-write head, leading to another potential source of wear-out. These lasers are rated for 6,000 hours of use, or 3.2PB of continuous data transferred per head. This is 20× higher than the workload specification that drives be able to transfer 17 MB/s on average for 5 years [6].

To deploy these denser drives, we need to reduce IO, but this is difficult. Storage systems already deploy large caches to take advantage of most locality in the storage accesses [36, 98, 136, 139]. Additional caching capacity also needs to be weighed against the cache's emissions. Caching also does not help with low-locality workloads, like LSM compaction [12, 43, 47, 48, 53, 90, 114]. Thus, we need to develop new solutions to reduce IO so we can deploy fewer, denser drives and reduce emissions.

Archival storage media. If we push using fewer, denser devices to the extreme, we need to consider media typically meant for archival storage: tape [117], glass [29], and DNA [52, 104]. All of these media have much longer access times, so we would need workloads that can tolerate these longer access times. The potential benefit is lower emissions. Tape has the potential to lower emissions by 87% per bit [80]. Unfortunately, this estimate does not include the robots and climate control needed to deploy tape, which significantly offsets its emissions reduction. Both glass and DNA are much denser than tape, so they have the potential to reduce emissions, but we cannot determine their emissions potential until more data is available on their lifecycle embodied and operational emissions, particularly when factoring in their achievable IO.

4.3 Extending lifetime

The last method to reducing storage emissions is extending device lifetime, which amortizes embodied emissions. Expected lifetime in servers has already increased from the traditional estimate of 3 years to 5-7 years, depending on the datacenter [25, 75, 96, 97].

Extending storage lifetime comes with drawbacks, some of which are shared with compute. Newer devices tend to be more energy-efficient, so in environments with significant operational emissions (e.g., with few renewables), extending lifetime can be detrimental to overall emissions. Embodied emissions are far more dominant in storage than in compute [70, 128], making this less of a concern. However, extending lifetime also has diminishing benefits. Going from 3 to 6 years halves embodied emissions, but halving again

requires going to 12 years. Meanwhile, failure rates increase with device age. It is thus not profitable to extend lifetime indefinitely.

Extending lifetime causes extra failures. Extending lifetimes in storage has the additional challenge of ensuring durability. As devices have longer lifetimes, their failure rate will increase. Although component failure is the expectation in datacenters, storage systems tune redundancy based on the likelihood of component failure. Additional failures require additional redundancy.

The benefit of extending lifetimes depends on how quickly failures increase for different device types. For HDDs, this is hard to predict — mostly because datacenters decomission HDDs before failure [83]. Reported annual failure rates can double going from three to six year lifetimes [83] as HDDs enter end of life [56, 57, 138]. Another challenge with hard disk drive lifetime is that we do not know the reasons for device failures. HDDs are assumed to fail from both wear-and-tear over the years and from IO utilization, but we do not have long-term studies of IO utilization to understand the significance of both of these factors.

For SSDs, extending lifetimes exacerbates flash's write endurance problem. Running the same workload on an SSD for double the lifetime doubles the writes. For workloads such as caching that already use most of the device's write endurance, this will likely cause the device to fail [99]. Together, extending lifetime and using denser flash to reduce emissions will require significant IO reduction.

Mitigating extra failures. Adaptive redundancy and enabling partial failures can mitigate these extra failures.

Adaptive redundancy was developed to enable lower capacity erasure-coding schemes during the useful life phase of HDD deployment [81–83]. For extending device lifetime, a similar idea could ensure durability at older ages — without requiring additional capacity overhead during the traditional lifetime. This reduction from embodied emissions will have to be balanced with transitioning the erasure codes with age, which causes additional IO that stresses bandwidth particularly for denser drives (Sec. 4.2).

Another way to mitigate the increased failure rates is to embrace partial failures. Although storage devices present a fixed capacity, this is not the reality. SSD cells wear out at different rates. HAMR HDDs can have some lasers fail. Sectors on HDDs can grow defects. While devices today can handle a limited number of defect failures, the device must fail if it no longer has the advertised fixed capacity. Thus, these partial failures are total failures today, causing us to lose usable capacity that we have already paid the embodied emissions for. We need to reconsider total failure and enable partial failure by changing the storage stack and how clouds deploy and replace drives. For HDDs particularly, while we generally know that annual failure rates increase with age [109, 116], we do not have the telemetry to know why exactly the device failed, limiting our ability to determine the emission benefits of partial failure. For instance, a HAMR drive with one laser failing results in a partial failure whereas the drive's only actuator no longer being reliable results in a complete drive failure since no part of the device is readable.

Second life and recycling. Another way to reduce embodied emissions is to increase hardware's lifetime through giving it a second life [70, 126] or recycling components [14]. Although giving storage drives a second life is more carbon-efficient [59], we would need to thoroughly address security concerns, especially if drives leave the datacenter [14, 59]. Second-life devices in the datacenter would need to be used for more failure-tolerant applications, such as caching.

5 A Call to Action

We identify three broad directions to reduce storage emissions, which each require significant further research. Much of the required research is interdisciplinary, requiring collaboration across the hardware/software boundary, across the entire software stack, as well as inputs from material scientists and sustainability analysts.

6 Acknowledgements

Sara McAllister is supported by a NDSEG Fellowship. We thank the members and companies of the PDL Consortium (Amazon, Google, Hitachi, Honda, IBM Research, Intel, Jane Street, Meta, Microsoft Research, Oracle, Pure Storage, Salesforce, Samsung, Two Sigma, Western Digital) for their interest, insights, feedback, and support. We thank our anonymous reviewers for their helpful comments and suggestions. We also thank Arie van der Hoeven and Praveen Viraraghavan at Seagate and Shruti Sethi at Microsoft for providing their technical expertise and suggestions.

References

[1] Disk prices. https://jcmit.net/diskprice.htm.

[2] Dv5 and Dsv5-series. https://learn.microsoft.com/en-us/azure/virtual-machines/dv5-dsv5-series. (Accessed on 05/09/2024).

[3] EBS default volume type updated to GP2. https://aws.amazon.com/about-aws/whats-new/2019/07/ebs-default-volume-type-updated-to-gp2/. (Accessed on 04/26/2024).

[4] Flash prices. https://jcmit.net/flashprice.htm.

[5] General Purpose Instance Types. https://aws.amazon.com/ec2/instance-types/. (Accessed on 05/09/2024).

[6] HAMR Reliability Tests Exceed Industry Standards. https://www.seagate.com/blog/mach2-and-hamr-breakthrough-ocp/. (Accessed on 05/09/2024).

[7] Helping you pick the greenest region for your Google Cloud resources. https://cloud.google.com/blog/topics/sustainability/pick-the-google-cloud-region-with-the-lowest-co2. (Accessed on 04/26/2024).

[8] IMEC netzero virtual fab. https://netzero.imec-int.com/. (Accessed on 04/26/2024).

[9] Is there a limit to the number of layers in 3d-nand? https://semiengineering.com/is-there-a-limit-to-the-number-of-layers-in-3d-nand/.

[10] Machine families resource and comparison. https://cloud.google.com/compute/docs/machine-resource. (Accessed on 05/09/2024).

[11] Makersite Data Platform. https://makersite.io/makersite-ai-data-apps/. (Accessed on 04/26/2024).

[12] Rocksdb. http://rocksdb.org.

[13] Samsung plans big capacity jump for ssds, preps 290-layer v-nand this year, 430-layer for 2025. https://www.tomshardware.com/pc-components/ssds/samsung-plans-big-capacity-jump-for-ssds-preps-290-layer-v-nand-this-year-430-layer-for-2025.

[14] Value recovery project, phase 2. https://thor.inemi.org/webdownload/2019/iNEMI-Value_Recovery2_Report.pdf.

[15] Wd and tosh talk up penta-level cell flash. https://blocksandfiles.com/2019/08/07/penta-level-cell-flash/ 5/17/22.

[16] Skyhawk datasheet. https://www.seagate.com/www-content/datasheets/pdfs/skyhawk-3-5-hddDS1902-6-1710US-en_US.pdf, 2017.

[17] Exos x10 data sheet. https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x-10DS1948-1-1709-GB-en_GB.pdf, 2021.

[18] Exos x18 data sheet. https://www.seagate.com/content/dam/seagate/migrated-assets/www-content/datasheets/pdfs/exos-x18-channel-DS2045-4-2106US-en_US.pdf, 2021.

[19] Exos 2x18 data sheet. https://www.seagate.com/content/dam/seagate/migrated-assets/www-content/datasheets/pdfs/exos-2x18-DS2093-1-2202US-en_US.pdf, 2022.

[20] Exos x10 sustainability report. https://www.seagate.com/content/dam/seagate/assets/esg/planet/product-sustainability/images/exos-x10-10tb-sustainability-report-2022/files/exos-x10-10tb.pdf, 2022.

[21] Exos x18 sustainability report. https://www.seagate.com/content/dam/seagate/assets/esg/planet/product-sustainability/images/exos-x18-sustainability-report/files/Exos-X18-18TB-Sustainability-Report-2023.pdf, 2022.

[22] Micron 7450 ssd with nvme. https://www.micron.com/content/dam/micron/global/public/products/product-flyer/7450-nvme-ssd-product-brief.pdf, 2022.

[23] Azure premium storage: Design for high performance. https://learn.microsoft.com/en-us/azure/virtual-machines/premium-storage-performance, 2023.

[24] Heat assisted marnetic recording HAMR. https://www.seagate.com/innovation/hamr/, 2024.

[25] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

[26] Nitin Agrawal, Vijayan Prabhakaran, Ted Wobber, John D. Davis, Mark Manasse, and Rina Panigrahy. Design tradeoffs for ssd performance. In *USENIX 2008 Annual Technical Conference*, ATC'08, page 57–70, USA, 2008. USENIX Association.

[27] Ahmed Amer, JoAnne Holliday, Darrell DE Long, Ethan L Miller, Jehan-François Pâris, and Thomas Schwarz. Data management and layout for shingled magnetic recording. *IEEE Transactions on Magnetics*, 47(10):3691–3697, 2011.

[28] Hrishikesh Amur, James Cipar, Varun Gupta, Gregory R. Ganger, Michael A. Kozuch, and Karsten Schwan. Robust and flexible power-proportional storage. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, page 217–228. Association for Computing Machinery, 2010.

[29] Patrick Anderson, Erika Blancada Aranas, Youssef Assaf, Raphael Behrendt, Richard Black, Marco Caballero, Pashmina Cameron, Burcu Canakci, Thales De Carvalho, Andromachi Chatzieleftheriou, Rebekah Storan Clarke, James Clegg, Daniel Cletheroe, Bridgette Cooper, Tim Deegan, Austin Donnelly, Rokas Drevinskas, Alexander Gaunt, Christos Gkantsidis, Ariel Gomez Diaz, Istvan Haller, Freddie Hong, Teodora Ilieva, Shashidhar Joshi, Russell Joyce, Mint Kunkel, David Lara, Sergey Legtchenko, Fanglin Linda Liu, Bruno Magalhaes, Alana Marzoev, Marvin Mcnett, Jayashree Mohan, Michael Myrah, Trong Nguyen, Sebastian Nowozin, Aaron Ogus, Hiske Overweg, Antony Rowstron, Maneesh Sah, Masaaki Sakakura, Peter Scholtz, Nina Schreiner, Omer Sella, Adam Smith, Ioan Stefanovici, David Sweeney, Benn Thomsen, Govert Verkes, Phil Wainman, Jonathan Westcott, Luke Weston, Charles Whittaker, Pablo Wilke Berenguer, Hugh Williams, Thomas Winkler, and Stefan Winzeck. Project silica: Towards sustainable cloud archival storage in glass. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, 2023.

[30] Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau. *Operating Systems: Three Easy Pieces*. Arpaci-Dusseau Books, 1.10 edition, November 2023.

[31] Desire Athow. Seagate launches biggest hard drive ever — 30tb exos mozaic 3+ hdd can store more than 1,000 blu-ray movies and, yes, everyone will be able to buy them. https://www.techradar.com/pro/seagate-launches-biggest-hard-drive-ever-30tb-exos-mozaic-3-hdd-can-store-more-than-1000-blu-ray-movies-and-yes-everyone-will-be-able-to-buy-them, 2024.

[32] Backblaze. Erasure coding used by Backblaze. https://www.backblaze.com/blog/reed-solomon/, 2013-2018.

[33] 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.

[34] M Garcia Bardon, P Wuytens, L-Å Ragnarsson, G Mirabelli, D Jang, G Willems, A Mallik, A Spessot, J Ryckaert, and B Parvais. Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 41–4. IEEE, 2020.

[35] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. Enabling Sustainable Clouds: The Case for Virtualizing the Energy System. In *Symposium on Cloud Computing*, 2021.

[36] Benjamin Berg, Daniel S. Berger, Sara McAllister, Isaac Grosof, Sathya Gunasekar, Jimmy Lu, Michael Uhlar, Jim Carrig, Nathan Beckmann, Mor Harchol-Balter, and Gregory G. Ganger. The CacheLib caching engine: Design and experiences at scale. In *USENIX OSDI*, 2020.

[37] Matias Bjørling, Abutalib Aghayev, Hans Holmberg, Aravind Ramesh, Damien Le Moal, Gregory R. Ganger, and George Amvrosiadis. ZNS: Avoiding the block interface tax for flash-based SSDs. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 689–703. USENIX Association, July 2021.

[38] Simona Boboila and Peter Desnoyers. Write endurance in flash drives: Measurements and analysis. In *USENIX FAST*, 2010.

[39] Erik Brunvand, Donald Kline, and Alex K. Jones. Dark silicon considered harmful: A case for truly green computing. In *2018 Ninth International Green and Sustainable Computing Conference (IGSC)*, 2018.

[40] Brad Calder, Ju Wang, Aaron Ogus, Niranjan Nilakantan, Arild Skjolsvold, Sam McKelvie, Yikang Xu, Shashwat Srivastav, Jiesheng Wu, Huseyin Simitci, et al. Windows azure storage: a highly available cloud storage service with strong consistency. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, pages 143–157, 2011.

[41] Enrique V. Carrera, Eduardo Pinheiro, and Ricardo Bianchini. Conserving disk energy in network servers. In *Proceedings of the 17th Annual International Conference on Supercomputing*, ICS '03, page 86–97, New York, NY, USA, 2003. Association for Computing Machinery.

[42] John A. Chandy. Dual actuator logging disk architecture and modeling. *Journal of Systems Architecture*, 53(12):913–926, 2007.

[43] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.

[44] Gong Chen, Wenbo He, Jie Liu, Suman Nath, Leonidas Rigas, Lin Xiao, and Feng Zhao. Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services. In *Symposium on Networked Systems Design and Implementation*, 2008.

[45] Shuang Chen, Christina Delimitrou, and José F Martínez. PARTIES: QoS-Aware Resource Partitioning for Multiple Interactive Services. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019.

[46] D. Colarelli and D. Grunwald. Massive arrays of idle disks for storage archives. In *SC '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, pages 47–47, 2002.

[47] Niv Dayan, Manos Athanassoulis, and Stratos Idreos. Monkey: Optimal navigable key-value store. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, 2017.

[48] Niv Dayan and Stratos Idreos. Dostoevsky: Better space-time trade-offs for lsm-tree based key-value stores via adaptive removal of superfluous merging. In *Proceedings of the 2018 International Conference on Management of Data*,

SIGMOD '18, 2018.
[49] Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56:74–80, 2013.
[50] Christina Delimitrou and Christos Kozyrakis. Paragon: QoS-aware Scheduling for Heterogeneous Datacenters. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2013.
[51] Christina Delimitrou and Christos Kozyrakis. Quasar: Resource-Efficient and QoS-Aware Cluster Management. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2014.
[52] George Dickinson, Golam Mortuza, William Clay, Luca Piantanida, Christopher Green, Chad Watson, Eric Hayden, Tim Andersen, Wan Kuang, Elton Graugnard, and William Hughes. An alternative approach to nucleic acid memory. *Nature Communications*, 12, 04 2021.
[53] Siying Dong, Shiva Shankar P, Satadru Pan, Anand Ananthabhotla, Dhanabal Ekambaram, Abhinav Sharma, Shobhit Dayal, Nishant Vinaybhai Parikh, Yanqin Jin, Albert Kim, Sushil Patil, Jay Zhuang, Sam Dunster, Akanksha Mahajan, Anirudh Chelluri, Chaitanya Datye, Lucas Vasconcelos Santana, Nitin Garg, and Omkar Gawde. Disaggregating rocksdb: A production experience. *Proc. ACM Manag. Data*, 2023.
[54] Lieven Eeckhout. Focal: A first-order carbon model to assess processor sustainability. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 401–415, 2024.
[55] Assaf Eisenman, Asaf Cidon, Evgenya Pergament, Or Haimovich, Ryan Stutsman, Mohammad Alizadeh, and Sachin Katti. Flashield: a hybrid key-value cache that controls flash write amplification. In *USENIX NSDI*, 2019.
[56] J.G. Elerath. Afr: problems of definition, calculation and measurement in a commercial environment. In *Annual Reliability and Maintainability Symposium. 2000 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.00CH37055)*, pages 71–76, 2000.
[57] J.G. Elerath. Specifying reliability in the disk drive industry: No more mtbf's. In *Annual Reliability and Maintainability Symposium. 2000 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.00CH37055)*, pages 194–199, 2000.
[58] Apache Software Foundation. HDFS Erasure Coding. https://hadoop.apache.org/docs/r3.0.0/hadoop-project-dist/hadoop-hdfs/HDFSErasureCoding.html, 2017 (accessed September 23, 2019).
[59] Kali Frost, Ines Sousa, Joanne Larson, Hongyue Jin, and Inez Hua. Environmental impacts of a circular recovery process for hard disk drive rare earth magnets. *Resources, Conservation and Recycling*, 173:105694, 2021.
[60] Kaihua Fu, Wei Zhang, Quan Chen, Deze Zeng, Xin Peng, Wenli Zheng, and Minyi Guo. QoS-Aware and Resource Efficient Microservice Deployment in Cloud-Edge Continuum. In *International Parallel and Distributed Processing Symposium*, 2021.
[61] Yu Gan, Mingyu Liang, Sundar Dev, David Lo, and Christina Delimitrou. Sage: Practical and Scalable ML-Driven Performance Debugging in Microservices. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2021.
[62] Yu Gan, Yanqi Zhang, Dailun Cheng, Ankitha Shetty, Priyal Rathi, Nayan Katarki, Ariana Bruno, Justin Hu, Brian Ritchken, Brendon Jackson, Kelvin Hu, Meghna Pancholi, Yuan He, Brett Clancy, Chris Colen, Fukang Wen, Catherine Leung, Siyuan Wang, Leon Zaruvinsky, Mateo Espinosa, Rick Lin, Zhongling Liu, Jake Padilla, and Christina Delimitrou. An Open-Source Benchmark Suite for Microservices and Their Hardware-Software Implications for Cloud & Edge Systems. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019.
[63] Anshul Gandhi, Mor Harchol-Balter, and Michael A. Kozuch. The case for sleep states in servers. In *Proceedings of the 4th Workshop on Power-Aware Computing and Systems*, HotPower '11, New York, NY, USA, 2011. Association for Computing Machinery.
[64] Jiechao Gao, Haoyu Wang, and Haiying Shen. Smartly Handling Renewable Energy Instability in Supporting A Cloud Datacenter. In *International Parallel and Distributed Processing Symposium*, 2020.
[65] Yixiao Gao, Qiang Li, Lingbo Tang, Yongqing Xi, Pengcheng Zhang, Wenwen Peng, Bo Li, Yaohui Wu, Shaozong Liu, Lei Yan, et al. When cloud storage meets {RDMA}. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 519–533, 2021.
[66] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 29–43, 2003.
[67] Garth R Goodson, Jay J Wylie, Gregory R Ganger, and Michael K Reiter. Efficient byzantine-tolerant erasure-coded storage. In *International Conference on Dependable Systems and Networks, 2004*, pages 135–144. IEEE, 2004.
[68] P.M. Greenawalt. Modeling power management for hard disks. In *Proceedings of International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 62–66, 1994.
[69] E. Grochowski and D.A. Thompson. Outlook for maintaining areal density growth in magnetic recording. *IEEE Transactions on Magnetics*, 30(6):3797–3800, 1994.
[70] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. ACT: designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. ACM, 2022.
[71] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 854–867. IEEE, 2021.
[72] Sudhanva Gurumurthi, Anand Sivasubramaniam, Mahmut Kandemir, and Hubertus Franke. Drpm: dynamic speed control for power management in server class disks. In *Proceedings of the 30th Annual International Symposium on Computer Architecture*, ISCA '03, page 169–181, New York, NY, USA, 2003. Association for Computing Machinery.
[73] Jun He, Sudarsun Kannan, Andrea C Arpaci-Dusseau, and Remzi H Arpaci-Dusseau. The unwritten contract of solid state drives. In *ACM EuroSys*, 2017.
[74] Bruce Hoch and Sage Shih. Open Cloud Server - Project Olympus JBOD. http://files.opencompute.org/oc/public.php?service=files&t=ea8af1772e9eea08a0fc0f8e1691418b, 2017.
[75] Amy Hood. Microsoft earnings release fy22 q4. https://www.microsoft.com/en-us/Investor/earnings/FY-2022-Q4/press-release-webcast, 2022. (Accessed on 04/26/2024).
[76] Cheng Huang, Huseyin Simitci, Yikang Xu, Aaron Ogus, Brad Calder, Parikshit Gopalan, Jin Li, Sergey Yekhanin, et al. Erasure Coding in Windows Azure Storage. 2012.
[77] Jiao Hui, Xiongzi Ge, Xiaoxia Huang, Yi Liu, and Qiangjun Ran. E-hash: an energy-efficient hybrid storage system composed of one ssd and multiple hdds. In *Proceedings of the Third International Conference on Advances in Swarm Intelligence - Volume Part II*, ICSI'12, page 527–534, Berlin, Heidelberg, 2012. Springer-Verlag.
[78] Shixin Ji, Zhuoping Yang, Xingzhen Chen, Stephen Cahoon, Jingtong Hu, Yiyu Shi, Alex K. Jones, and Peipei Zhou. Scarif: Towards carbon modeling of cloud servers with accelerators, 2024.
[79] Zhipeng Jia and Emmett Witchel. Nightcore: Efficient and Scalable Serverless Computing for Latency-Sensitive, Interactive Microservices. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2021.
[80] Brad Johns. Reducing data center energy consumption and carbon emissions with modern tape storage. https://datastorage-na.fujifilm.com/wp-content/themes/fuji/images/sustainability/BJC-Reducing-Carbon-Emission-Whitepaper-LR-1120.pdf, 2020.
[81] Saurabh Kadekodi, Francisco Maturana, Sanjith Athlur, Arif Merchant, K. V. Rashmi, and Gregory R. Ganger. Tiger: Disk-Adaptive redundancy without placement restrictions. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 413–429, Carlsbad, CA, July 2022. USENIX Association.
[82] Saurabh Kadekodi, Francisco Maturana, Suhas Jayaram Subramanya, Juncheng Yang, KV Rashmi, and Gregory Ganger. PACEMAKER: Avoiding HeART attacks in storage clusters with disk-adaptive redundancy. In *Symposium on Operating Systems Design and Implementation*, 2020.
[83] Saurabh Kadekodi, K. V. Rashmi, and Gregory R. Ganger. Cluster storage systems gotta have HeART: improving storage efficiency by exploiting disk-reliability heterogeneity. In *17th USENIX Conference on File and Storage Technologies (FAST 19)*, pages 345–358, Boston, MA, February 2019. USENIX Association.
[84] Saurabh Kadekodi, Shashwat Silas, David Clausen, and Arif Merchant. Practical design considerations for wide locally recoverable codes (LRCs). In *21st USENIX Conference on File and Storage Technologies (FAST 23)*, pages 1–16, Santa Clara, CA, February 2023. USENIX Association.
[85] Rainer W. Kaese. From 20 megabytes to 20 terabytes: 40 years of hard disk drive technology. https://www.toshiba-storage.com/wp-content/uploads/2023/02/Toshiba_40years_HDD_technology_screen.pdf, 2022.
[86] William Katsak, Íñigo Goiri, Ricardo Bianchini, and Thu D. Nguyen. Greencassandra: Using renewable energy in distributed structured storage systems. In *2015 Sixth International Green and Sustainable Computing Conference (IGSC)*, pages 1–8, 2015.
[87] Rini T. Kaushik and Milind Bhandarkar. Greenhdfs: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster. In *Proceedings of the 2010 International Conference on Power Aware Computing and Systems*, HotPower'10, page 1–9, USA, 2010. USENIX Association.
[88] M. Kobayashi, S. Nakagawa, and H. Numasato. Adaptive control of dual-stage actuator for hard disk drives. In *Proceedings of the 2004 American Control Conference*, volume 1, pages 523–528 vol.1, 2004.
[89] Changman Lee, Dongho Sim, Jooyoung Hwang, and Sangyeun Cho. F2fs: A new file system for flash storage. In *USENIX FAST*, 2015.
[90] Baptiste Lepers, Oana Balmau, Karan Gupta, and Willy Zwaenepoel. Kvell: The design and implementation of a fast persistent key-value store. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, SOSP '19, 2019.

[91] Cheng Li, Philip Shilane, Fred Douglis, and Grant Wallace. Pannier: Design and analysis of a container-based flash cache for compound objects. *ACM Transactions on Storage*, 13(3):24, 2017.

[92] Daping Li, Xiaoyang Qu, Jiguang Wan, Jun Wang, Yang Xia, Xiaozhao Zhuang, and Changsheng Xie. Workload scheduling for massive storage systems with arbitrary renewable supply. *IEEE Transactions on Parallel and Distributed Systems*, 29(10):2373–2387, 2018.

[93] Lanyue Lu, Thanumalayan Sankaranarayana Pillai, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Wisckey: Separating keys from values in ssd-conscious storage. In *USENIX FAST*, 2016.

[94] Shutian Luo, Huanle Xu, Chengzhi Lu, Kejiang Ye, Guoyao Xu, Liping Zhang, Yu Ding, Jian He, and Chengzhong Xu. Characterizing Microservice Dependency and Performance: Alibaba Trace Analysis. In *Symposium on Cloud Computing*, 2021.

[95] Shutian Luo, Huanle Xu, Kejiang Ye, Guoyao Xu, Liping Zhang, Jian He, Guodong Yang, and Chengzhong Xu. Erms: Efficient Resource Management for Shared Microservices with SLA Guarantees. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2022.

[96] Jialun Lyu, Jaylen Wang, Kali Frost, Chaojie Zhang, Celine Irvene, Esha Choukse, Rodrigo Fonseca, Ricardo Bianchini, Fiodar Kazhamiaka, and Daniel S. Berger. Myths and misconceptions around reducing carbon embedded in cloud platforms. In *2nd Workshop on Sustainable Computer Systems (HotCarbon23)*. ACM, July 2023.

[97] Jialun Lyu, Marisa You, Celine Irvene, Mark Jung, Tyler Narmore, Luke Shapiro, Jacob an d Marshall, Savyasachi Samal, Ioannis Manousakis, Lisa Hsu, et al. Hyrax:{Fail-in-Place} server operation in cloud platforms. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 287–304, 2023.

[98] Sara McAllister, Benjamin Berg, Julian Tutuncu-Macias, Juncheng Yang, Sathya Gunasekar, Jimmy Lu, Daniel S. Berger, Nathan Beckmann, and Gregory R. Ganger. Kangaroo: Caching billions of tiny objects on flash. In *ACM SOSP*, 2021.

[99] Sara McAllister, Yucong Wang, Benjamin Berg, Daniel S. Berger, Nathan Beckmann, and Gregory R. Ganger. Fairywren: A sustainable cache for emerging write-read-erase flash interfaces. In *USENIX OSDI*, 2024.

[100] Jaehong Min, Chenxingyu Zhao, Ming Liu, and Arvind Krishnamurthy. eZNS: An elastic zoned namespace for commodity ZNS SSDs. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 461–477, Boston, MA, July 2023. USENIX Association.

[101] Amirhossein Mirhosseini, Sameh Elnikety, and Thomas F Wenisch. Parslo: A Gradient Descent-based Approach for Near-optimal Partial SLO Allotment in Microservices. In *Symposium on Cloud Computing*, 2021.

[102] S.W. Ng. Improving disk performance via latency reduction. *IEEE Transactions on Computers*, 40(1):22–30, 1991.

[103] S.W. Ng. Latency reduction for cd-rom and clv disks. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, volume i, pages 100–108 vol.1, 1992.

[104] Bichlien Nguyen, Julie Sinistore, Jake Smith, Praneet S. Arshi, Lauren M. Johnson, Tim Kidman, T.J. diCaprio, Doug Carmean, and Karin Strauss. Architecting datacenters for sustainability: Greener data storage using synthetic dna. In *Electronics Goes Green 2020*. Fraunhofer IZM, IEEE, September 2020.

[105] Satadru Pan, Theano Stavrinos, Yunqiao Zhang, Atul Sikaria, Pavel Zakharov, Abhinav Sharma, Mike Shuey, Richard Wareing, Monika Gangapuram, Guanglei Cao, et al. Facebook's tectonic filesystem: Efficiency from exascale. In *19th USENIX Conference on File and Storage Technologies (FAST 21)*, pages 217–231, 2021.

[106] David A. Patterson, Garth Gibson, and Randy H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). 1988.

[107] Eduardo Pinheiro and Ricardo Bianchini. Energy conservation techniques for disk array-based servers. In *Proceedings of the 18th Annual International Conference on Supercomputing*, ICS '04, page 68–78, New York, NY, USA, 2004. Association for Computing Machinery.

[108] Eduardo Pinheiro, Ricardo Bianchini, and Cezary Dubnicki. Exploiting redundancy to conserve energy in storage systems. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '06/Performance '06, page 15–26, New York, NY, USA, 2006. Association for Computing Machinery.

[109] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. Failure Trends in a Large Disk Drive Population. In *Conference on File and Storage Technologies*, 2007.

[110] Francisco Pires. Solidigm introduces industry-first plc nand for higher storage densities. https://www.tomshardware.com/news/solidigm-plc-nand-ssd, 2022.

[111] Haoran Qiu, Subho S Banerjee, Saurabh Jha, Zbigniew T Kalbarczyk, and Ravishankar K Iyer. FIRM: An Intelligent Fine-grained Resource Management Framework for SLO-Oriented Microservices. In *Symposium on Operating Systems Design and Implementation*, 2020.

[112] Xiaoyang Qu, Jiguang Wan, Jun Wang, Liqiong Liu, Dan Luo, and Changsheng Xie. Greenmatch: Renewable-aware workload scheduling for massive storage systems. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 403–412, 2016.

[113] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, and Nick Care. Carbon-Aware Computing for Datacenters. *Transactions on Power Systems*, 38:1270–1280, 2022.

[114] Pandian Raju, Rohan Kadekodi, Vijay Chidambaram, and Ittai Abraham. Pebblesdb: Building key-value stores using fragmented log-structured merge trees. In *ACM SOSP*, 2017.

[115] K V Rashmi, Nihar B Shah, Dikang Gu, Hairong Kuang, Dhruba Borthakur, and Kannan Ramchandran. A hitchhiker's guide to fast and efficient data reconstruction in erasure-coded data centers. 2014.

[116] Bianca Schroeder and Garth A. Gibson. Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In *5th USENIX Conference on File and Storage Technologies (FAST 07)*, San Jose, CA, February 2007. USENIX Association.

[117] T.J.E. Schwarz, Qin Xin, E.L. Miller, D.D.E. Long, A. Hospodor, and S. Ng. Disk scrubbing in large archival storage systems. In *The IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, 2004. (MASCOTS 2004). Proceedings.*, pages 409–418, 2004.

[118] Seagate. The Digitization of the World From Edge to Core. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, 2018.

[119] Mark A. Shaw. Project Olympus Flash Expansion FX-16. http://files.opencompute.org/oc/public.php?service=files&t=14ab3cf25170b7a0a439e11a3d818c96, 2017.

[120] Anton Shilov. Seagate reveals hamr hdd roadmap: 32tb first, 40tb follows. https://www.tomshardware.com/news/seagate-reveals-hamr-roadmap-32-tb-comes-first, 2023.

[121] Junaid Shuja, Kashif Bilal, Sajjad A. Madani, Mazliza Othman, Rajiv Ranjan, Pavan Balaji, and Samee U. Khan. Survey of Techniques and Architectures for Designing Energy-Efficient Data Centers. *IEEE Systems Journal*, 10:507–519, 2016.

[122] Minseok Song, Yeongju Lee, and Euiseok Kim. Saving disk energy in video servers by combining caching and prefetching. *ACM Trans. Multimedia Comput. Commun. Appl.*, 10(1s), jan 2014.

[123] Abel Souza, Noman Bashir, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

[124] Akshitha Sriraman and Thomas F Wenisch. μTune: Auto-Tuned Threading for OLDI Microservices. In *Conference on Operating Systems Design and Implementation*, 2018.

[125] Chetan Choppali Sudarshan, Nikhil Matkar, Sarma Vrudhula, Sachin S Sapatnekar, and Vidya A Chhabria. Eco-chip: Estimation of carbon footprint of chiplet-based architectures for sustainable vlsi. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 671–685. IEEE, 2024.

[126] Jennifer Switzer, Gabriel Marcano, Ryan Kastner, and Pat Pannuto. Junkyard Computing: Repurposing Discarded Smartphones to Minimize Carbon. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023.

[127] Linpeng Tang, Qi Huang, Wyatt Lloyd, Sanjeev Kumar, and Kai Li. RIPQ: advanced photo caching on flash for facebook. In *USENIX FAST*, 2015.

[128] Swamit Tannu and Prashant J Nair. The Dirty Secret of SSDs: Embodied Carbon. In *HotCarbon*, 2022.

[129] Eno Thereska, Austin Donnelly, and Dushyanth Narayanan. Sierra: practical power-proportionality for data center storage. In *Proceedings of the Sixth Conference on Computer Systems*, EuroSys '11, page 169–182, New York, NY, USA, 2011. Association for Computing Machinery.

[130] Amanda Tomlinson and George Porter. Something Old, Something New: Extending the Life of CPUs in Datacenters. In *HotCarbon*, 2022.

[131] Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, and Akshitha Sriraman. Designing cloud servers for lower carbon. In *ISCA*, 2024.

[132] Jaylen Wang, Udit Gupta, and Akshitha Sriraman. Characterizing Datacenter Server Generations for Lifetime Extension and Carbon Reduction. In *Workshop on NetZero Carbon Computing*, 2023.

[133] Jaylen Wang, Udit Gupta, and Akshitha Sriraman. Giving Old Servers New Life at Hyperscale. In *Workshop on Hot Topics in System Infrastructure*, 2023.

[134] Jaylen Wang, Udit Gupta, and Akshitha Sriraman. Peeling Back the Carbon Curtain: Carbon Optimization Challenges in Cloud Computing. In *Workshop on Sustainable Computer Systems*, 2023.

[135] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In *International Middleware Conference*, 2021.

[136] Daniel Lin-Kit Wong, Hao Wu, Carson Molder, Sathya Gunasekar, Jimmy Lu, Snehal Khandkar, Abhinav Sharma, Daniel S Berger, Nathan Beckmann, and Gregory R Ganger. Baleen:{ML} admission & prefetching for flash caches. In *22nd USENIX Conference on File and Storage Technologies (FAST 24)*, pages 347–371, 2024.

[137] Roger Wood, Mason Williams, Aleksandar Kavcic, and Jim Miles. The feasibility of magnetic recording at 10 terabits per square inch on conventional media. *IEEE Transactions on Magnetics*, 45(2):917–923, 2009.

[138] J. Yang and Feng-Bin Sun. A comprehensive review of hard-disk drive reliability. In *Annual Reliability and Maintainability. Symposium. 1999 Proceedings (Cat. No.99CH36283)*, pages 403–409, 1999.

[139] Juncheng Yang, Yao Yue, and Rashmi Vinayak. A large scale analysis of hundreds of in-memory cache clusters at twitter. In *USENIX OSDI*, 2020.

[140] Cristian Zambelli, Rino Micheloni, and Piero Olivo. Reliability challenges in 3d nand flash memories. In *2019 IEEE 11th International Memory Workshop (IMW)*, pages 1–4, 2019.

[141] Yanqi Zhang, Weizhe Hua, Zhuangzhuang Zhou, G. Edward Suh, and Christina Delimitrou. Sinan: ML-Based and QoS-Aware Resource Management for Cloud Microservices. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2021.

[142] Zhe Zhang, Andrew Wang, Kai Zheng, G Uma Maheswara, and B Vinayakumar. Introduction to hdfs erasure coding in apache hadoop. *blog.cloudera.com*, 2015.

[143] Aviad Zuck, Donald Porter, and Dan Tsafrir. Degrading data to save the planet. In *Proceedings of the 19th Workshop on Hot Topics in Operating Systems*, HOTOS '23, 2023.