How Using Spare Capacity for Data Storage is Better for the Environment

A research paper identifying the factors and calculating the impact of data storage on carbon emissions.

Contents

Abstract

The global datasphere is large and growing rapidly. Storage and distribution of that data already represents a significant and growing contributor to global carbon. While much attention is given to the amount of electricity consumed in powering and cooling data storage devices, it is important to note that the carbon and overall environmental impact from the manufacture and transport of storage devices is often as big or bigger than their ongoing operation. Additionally there is evidence that the vast majority of drives are severely underutilized and could be put into operation for data storage. It takes almost no additional electricity or cooling to run a drive at near full capacity versus partial capacity. Thus, significant carbon savings can be achieved by enabling already powered and deployed drives to be more fully utilized during their lifetime.

This report also surfaces that additional significant carbon savings can be achieved by:

1. Increasing the effective life and utilization of already manufactured drives.
2. Minimizing the number of copies of data that need to be stored
    a. to achieve equivalent data durability levels.
    b. to achieve equivalent global geographic distribution and performance levels.
3. Enabling a carbon efficient distribution of storage between HDD and SSD.

Storj is a system that is designed to provide enterprise grade, S3 compatible cloud object storage by leveraging underutilized capacity from independently operated drives around the world. This research paper identifies the carbon costs of centralized cloud storage, evaluates the ways that Storj operates differently, and presents a model to estimate carbon footprint savings from using Storj as compared to hypercloud and traditional data center storage.

The model produces estimates of significant carbon savings of 66-83% per 1 TB of enterprise data over a three year period. This is equivalent to greater than 200 kg CO_2/TB.

Introduction

According to IDC, "in 2020, 64.2 zettabytes of data was created or replicated" and IDC forecasted that "global data creation and replication will experience a compound annual growth rate (CAGR) of 23% over the 2020-2025 forecast period." At that rate, more than 180 zettabytes — or 180 billion terabytes — will be created and replicated in 2025.[1]

> "Global data creation and replication will experience a compound annual growth rate (CAGR) of 23% over the 2020-2025 forecast period."
>
> IDC

Replication is a multiplying factor in the storage capacity needed for data growth. Given the importance of data durability, most organizations store multiple copies of a given set of data to protect against things like drive failure, accidental overwrite, fire, floods, etc. Copies are also made to provide access in various regions. Most hyperscalers store a minimum of three copies of data. The 3-2-1 Backup Rule (store 3 copies of data, in 2 different types of media, with at least 1 copy offsite) is a commonly cited best practice.[2]

If a company needs to store 3 copies of a given set of data, then that data has an expansion factor of 3. One functional TB of data ends up needing 3 TB of capacity, across 3 different drives, with the associated economic and carbon cost of those drives.

In order to achieve reasonable assurances of durability, most companies rely on replication. Most cloud storage providers replicate data 3X within a region in order to achieve sufficient durability. But, if the data is to be stored multi-region, for disaster recovery or for distribution purposes, that number is multiplied by the number of regions.

---

[1] Worldwide IDC Global DataSphere Forecast 2022-2026

[2] US Chamber 3-2-1 Backup Rule

Figure 1: Calculated file durability for various numbers of copies (on separate drives) and various projected drive failure rates.[3]

Durability (%)			
Copies	.5% annual failure rate	1.0% annual failure rate	2.5% annual failure rate
1	99.501247919268200	99.004983374916800	97.530991202833200
2	99.995033208665900	99.980264677289000	99.879089572574900
3	99.999944379031300	99.999560004506000	99.993352635849300
4	99.999999343911600	99.999989669042200	99.999615316607400
5	99.999999992029700	99.999999750204800	99.999977080897800
6	99.999999999901300	99.999999993844600	99.999998608598100
7	99.999999999998700	99.999999999846300	99.999999914411300
8	99.999999999999900	99.999999999996100	99.999999994683800
9	99.999999999999900	99.999999999999900	99.999999999667300

The number of copies that need to be stored using replication depends on the expected drive failure rate and the level of durability required. The table above shows that, if replication is used, and there is an expectation of 0.5% annual drive failure, then 4 copies of data need to be stored to achieve 8 9's of durability (indicated in yellow) and 6 copies need to be stored for 11 9's of durability (indicated in green). For drive failure rates of 1.0% and 2.5%, those numbers increase to 5 and 7 copies respectively for 8 9's of durability, and 7 and 9 copies for 11 9's of durability.

The combination of the natural growth of data generation and the need for replication for both backup and regional access causes a growth in hard drives that need to be manufactured. Additionally, this results in more data centers that will need to be built and powered. Both factors will add significantly to carbon emissions.

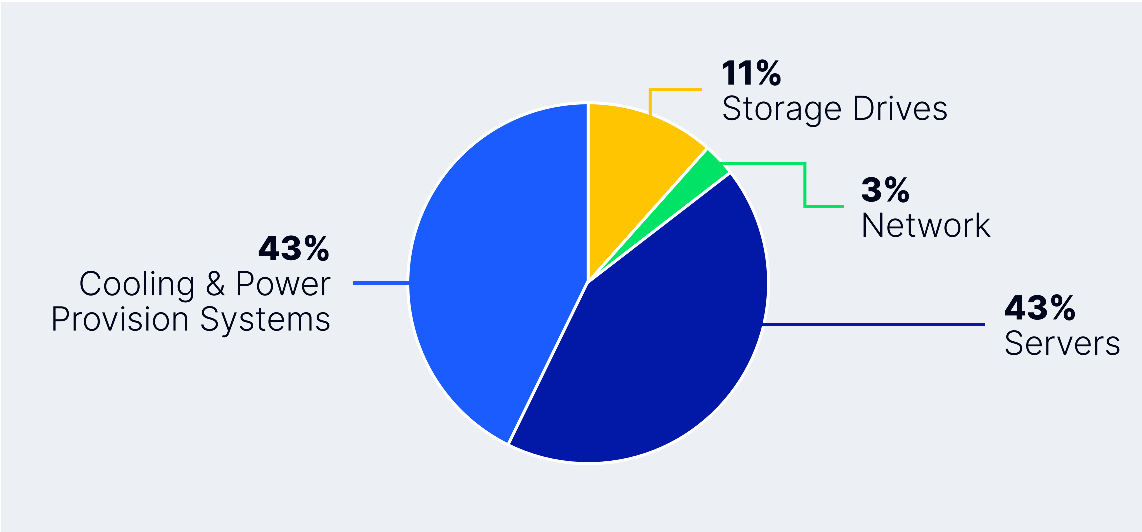
[3] Replication calculations for durability - See pages 21-22 of the Storj Decentralized Framework

Data storage factors

Factor #1: Energy needed to run data centers

The amount of power needed to run data centers on a global scale is estimated to be 416 terawatt-hours per year.[4] Although some sources place that number as high as 770 terawatt-hours per year.[5]

Figure 2: Fraction of U.S. data center electricity use by end use.[6]



Of that, it's estimated that 11% of the power is directly related to powering drives. And, some meaningful additional percentage of the energy consumed for cooling, power provisioning, compute, and network purposes is, of course, associated with the creation, transmission, management, and storage of data.[6]

Most hyperscalers can claim a great deal of carbon efficiency, because of their ability to amortize data center carbon overhead (plant, cooling, etc.) over large amounts of equipment, their ability to locate near cleaner sources of power, and their ability to efficiently utilize a large percentage of the capacity of equipment in the data center. In this sense, the large hyperscalers (e.g. AWS, Microsoft Azure, Google Cloud, Alibaba, IBM) may claim better carbon efficiency than smaller corporate, university, or tier two data centers. Because of this, and the fact that alternative storage methods still require energy, overhead factors were not used in the model to calculate reduction in carbon emissions.

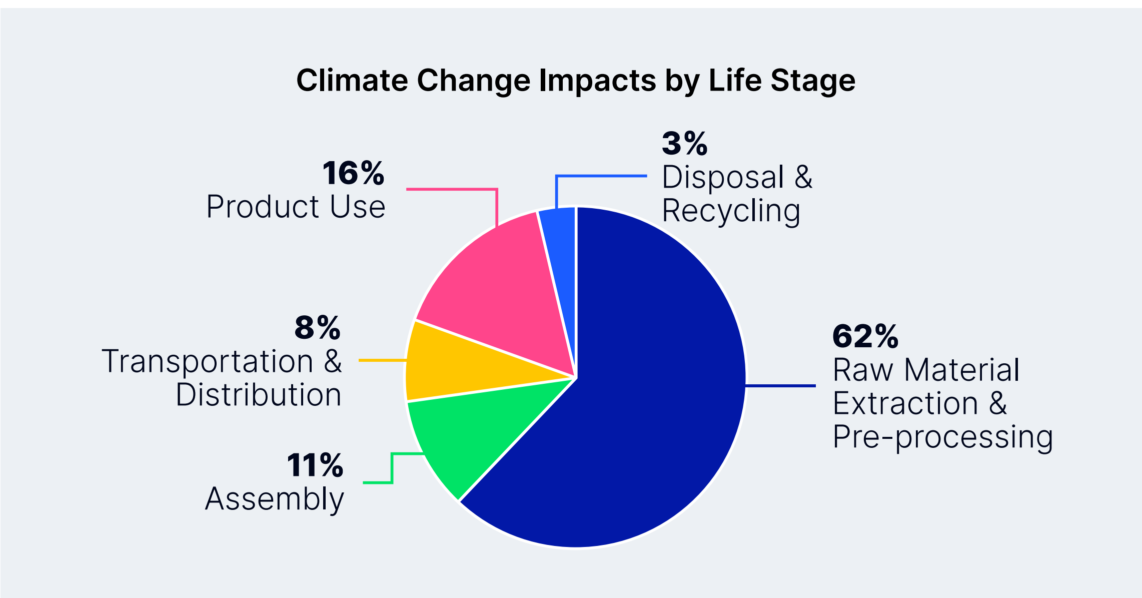
[4] C&C Tech Group - Understanding Data Center Energy Consumption
[5] IEA Data Centres and Data Transmission Networks Report
[6] Shehabi et. al. (2016). United States Data Center Energy Usage Report

That said, there are two other factors to note that are specific to data centers. One is the coolants used to keep the hard drives at safe operating temperatures. These are often made of hazardous chemicals and while not the impact was not readily quantifiable for this paper, disposal is a concern for the environment. A second factor is the batteries used as a backup if there are power shortages to try to maintain service. Both mining and disposal associated with batteries cause negative environmental impacts, but again were not quantifiable and therefore ignored in this model.

Factor #2: Manufacturing of storage drives

The environmental impact of data storage comes not only from powering drives and data centers, but also from the manufacturing, transport, and end of life phases of a drive's lifecycle. A significant portion of the embodied energy of a hard drive is derived from the manufacturing of the hard drive and takes place in various facilities across the globe.

Figure 3: Life cycle analysis of 1TB drive and its impact on the environment.[7]



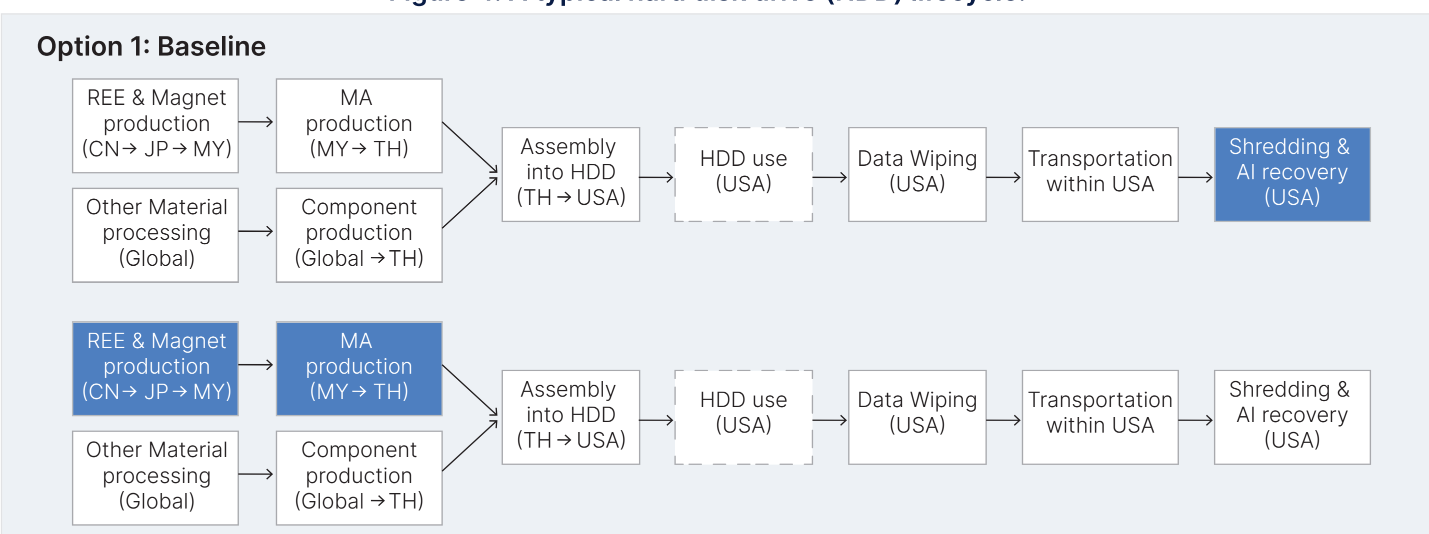
For example, the life cycle analysis (LCA) of a Seagate 1TB Momentus drive (shown above) demonstrates that only about 16% of the climate change impact comes from the energy used to power the drive. Raw Material Extraction and Pre-processing (62%), Assembly (11%), and Transportation & Distribution (8%) account for the bulk of the impact.[7]

[7] Seagate life cycle analysis of the Momentus drive and impact on climate change

A 2020 research paper published in Resources, Conservation, and Recycling,[8] discusses the manufacture of HDD in the context of various strategies for reusing and recycling drives. The manufacturing process begins with mining rare earth metals such as neodymium and dysprosium, primarily in China. The mined rare earth metals are then shipped to Japan to create magnets. The magnets are shipped to Malaysia for actuator assembly. The actuators are then shipped to Thailand for drive assembly.The drives are then shipped to their final destination. Authors noted the particularly devastating aspects of rare earth metal mining. “These impacts are especially concerning because it is estimated that 30% of global REEs are mined illegally (Packey and Kingsnorth, 2016), resulting in disastrous consequences for the environment and public health.”

A research report by Sabbaghi et al. on the global flow of hard drives noted that almost 1 billion HDDs reach end of life annually. The vast majority of these drives are wiped and shredded, with minimal recovery.[9] A report by Jin et al. on potential value recovery from end-of-life hard disk drives concluded that of various means to lessen the environmental impact of HDDs (reuse, recycling certain components, etc.), that reuse is the most environmentally impactful, saving approximately 5 kg CO_2, for every 6 months that a drive’s life is extended.[10]

Figure 4: A typical hard disk drive (HDD) lifecycle.[11]



[8] Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

[9] Mostafa Sabbaghi, et al, The Global Flow of Hard Disk Drives: Quantifying the Concept of Value Leakage in E-waste Recovery Systems.

[10] Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

[11] Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

Even for drives in constant use, the manufacture of the drives represents a huge component of the climate impact. Tannu and Nair in a report on embodied carbon in SSDs used the manufacturers' Life Cycle Analysis (LCA) of 26 drives of various sizes and manufacturers to estimate that the manufacture of 1 TB of HDD capacity results in a carbon footprint of about 20 kg of carbon. Powering that capacity over a five year lifetime results in another 76 kg of carbon.[12] But, even that 20 kg figure may understate the enormous cost, both in carbon and other environmental costs, from drive manufacture.

Referring to the research from Jin et al., the global warming impact of drive manufacture is but one of the environmental harms. The authors frame their research in terms of the savings from reuse of drives versus other forms of recycling (e.g. recycling of the actuators or magnets).

Figure 5: Savings from reuse of drives vs. other forms of recycling.[13]

Impact category	Unit	HDD reuse	MA reuse	M2M	Hydro	Pyro	ER
Ozone depletion	kg CFC-11 eq	3.7E-07	2.7E-07	1.1E-07	-4.2E-08	3.2E-07	-1.7E-08
Global Warming	kg CO_2 eq	5.5	1.9	0.7	-0.30	0.02	0.01
Smog	kg CO_3 eq	3.9E-01	1.7E-01	7.9E-02	8.5E-03	9.2E-02	2.4E-02
Acidification	kg SO_2 eq	4.6E-02	5.5E-02	3.1E-02	9.4E-03	3.4E-02	1.4E-02
Eutrophication	kg N eq	1.2E-02	2.2E-02	1.2E-02	6.7E-03	4.3E-02	7.7E-03
Carcinogenics	CTUh	3.1E-07	1.8E-07	3.4E-08	-4.0E-08	3.5E-07	2.4E-08
Non carcinogenics	CTUh	1.4E-06	4.5E-07	1.6E-07	6.9E-07	9.7E-06	7.6E-07
Respiratory effects	kg PM2.5 eq	5.2E-03	5.2E-03	2.7E-03	-3.6E-05	2.7E-03	5.0E-04
Ecotoxicity	CTUe	3.2E+01	1.0E+01	2.3E+00	1.4E+01	2.1E+02	1.7E+01
Fossil fuel depletion	MJ surplus	5.9E+00	2.4E+00	8.9E-01	-3.7E-01	6.0E-01	1.0E-01

The savings in the first column of Figure 5 represents what is achieved by extending the life of the drive for 6 months, and thus represents about 10% of the lifetime cost of drive manufacture. Not only does global warming from manufacturing one HDD represent 55 kg CO_2 in their model, but we also see a host of other significant harms during manufacture such as smog, eutrophication, carcinogens, and ecotoxicity.

[12] The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

[13] Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

Hypothesis

According to statistics from the NRDC and IBM, most servers operate at only 12-18% of capacity,[14] implying that there is a huge reservoir of already manufactured and powered, but severely underutilized, storage capacity. It is worth noting that it takes almost no additional electricity to run a drive at 80% capacity versus 20% capacity. It is worth noting that a disk spins and is powered the same amount whether it is running at 80% capacity or 20% capacity. Other than the nominal energy required for the initial write, an HDD drive will consume almost the exact same amount of energy at near full capacity as near empty.

> Therefore, one of the most impactful things we can do to reduce the carbon footprint of data is to ensure that drives that have already been produced are being used at near-full capacity.

If we could find a way to efficiently utilize spare capacity in drives that are already being spun, we can reduce the overall carbon footprint of storing and distributing data significantly. In addition to reusing HDDs, as Jin et. al. suggest, we can actually reduce the need to manufacture new drives in the first place by more efficiently storing data on existing drives. Using the Momentus drive example, by eliminating the need to manufacture and deploy a new drive, we save approximately 84% of the drive's lifetime carbon footprint created through raw material mining and extraction, assembly, transportation, and disposal. And, we eliminate a significant fraction of the remaining 16% associated with product use.

Therefore, one of the most impactful things we can do to reduce the carbon footprint of data is to ensure that drives that have already been produced are being used at near-full capacity. Further savings can be found by extending the life of already manufactured drives. By using drives at near capacity you reduce the need to manufacture new drives which have significant carbon footprints and use up rare materials. And with less drives needed, this also would reduce the need to build more data centers.

[14] Are Your Data Centers Keeping You From Sustainability?, IBM

Another area to consider for carbon reduction is the amount of data being stored. While entities will need to be as judicious as possible in what they deem necessary to store, data reduction is possible by reducing replication. This could be achieved by minimizing the expansion factor/number of copies of data that need to be stored to achieve equivalent data durability levels. Additional savings could be found by minimizing the expansion factor/number of copies of data that need to be stored to achieve equivalent global geographic distribution and performance levels.

> “It must yield both economic and carbon benefits, while delivering on the characteristics that matter in the data world, such as durability, availability, performance, security, and privacy. “
>
> Vinod Khosla

Lastly, because the manufacture of flash drives is nearly 8 times as carbon intensive[15] as HDDs, and their lifetime operation nearly twice as carbon intensive, this is another opportunity for improvement by enabling a carbon efficient distribution of storage between HDD and flash.

As Vinod Khosla noted in a presentation at Stanford on green tech, to be successful, green tech must first make economic sense.[16] To be more than an academic exercise, any system that meets the goals described above must not only provide significant green benefits, but also deliver tangible benefits to both the operators of drives and the consumers of data. It must yield both economic and carbon benefits, while delivering on the characteristics that matter in the data world, such as durability, availability, performance, security, and privacy.

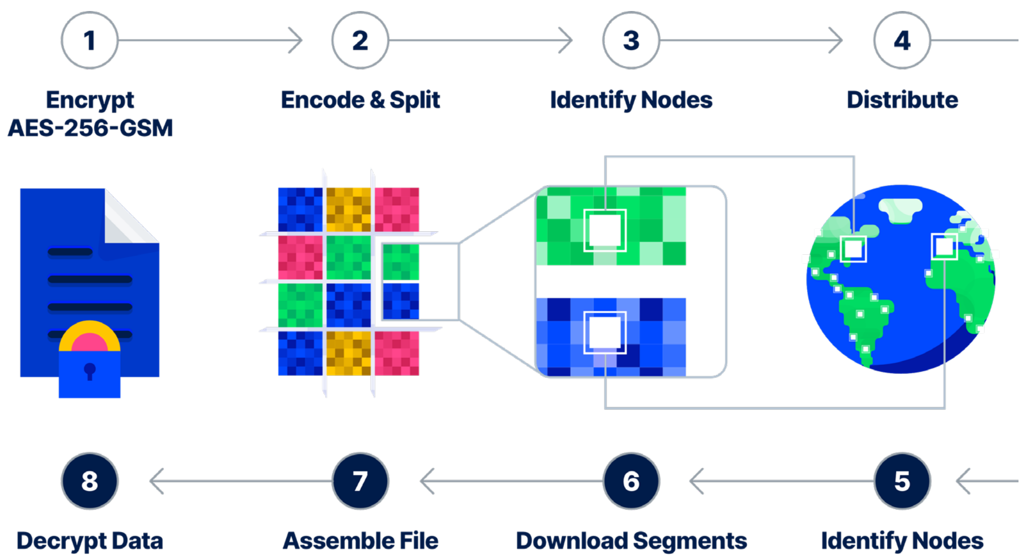
[15] The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

[16] Stanford Business: Vinod Khosla - Green Tech Must Make Economic Sense

A more sustainable alternative for data storage

Storj offers enterprise grade, S3 compatible cloud object storage, but does so by leveraging spare capacity from drives operated by individuals and data centers around the world. Those who rent out drives, called Storage Node Operators, or "SNOs", are compensated for the capacity and bandwidth that they contribute to the network. The Storj business model has been described as "Airbnb for Disk Drives."

Figure 6: How Storj is able to securely utilize spare storage capacity.



All data is encrypted and sharded using Reed Solomon erasure coding, and the resulting pieces are distributed around the world. A typical large file is divided into 64 MB encrypted segments, and each of those encrypted segments is divided into 80 erasure coded pieces, of which any 29 are needed to reconstitute the file. Each of those 80 pieces is then stored on a different drive around the world, across a multitude of different geographies, operators, power supplies, and networks. From a customer perspective, this allows Storj to deliver 11 9's of durability, superior global performance, and better security at about 1/10 to 1/20 the price of traditional cloud storage providers.

5 Ways Storj reduces carbon emissions

#1: Efficient use of underutilized storage capacity

Notably, all the steps that Storj takes to make storage secure, performant, durable, and economical also serve to make Storj a more efficient solution. Storj itself is not building out new data centers, purchasing drives, or consuming large amounts of electricity, etc. Thus, both the economic and direct carbon overhead is very modest. Furthermore, recent surveys of Storage Node Operators (SNOs) have concluded that about 21% of all capacity comes from drives that are already being run and powered. These SNOs experience no incremental costs from the purchase of drives and only marginal additional costs from operating those drives at higher capacity. Similarly, that additional capacity comes with no additional carbon footprint from drive manufacturing and only marginal additional carbon from operation.

Approximately 21% of the capacity in the network comes from older drives that were brought back online specifically for the purpose of becoming storage nodes. While the full amount of electricity used to operate those drives represents both an economic and carbon cost, the larger economic and carbon cost of drive manufacture and purchase are not seen. As a result, SNOs should be able to derive significant economic margin from drive operation, and Storj can sustainably both compensate SNOs fairly and offer storage to consumers at a fraction of the price of alternatives, while generating only modest additional carbon.

#2: Life extension of already manufactured drives

Given the carbon cost associated with drive manufacture and deployment and the economic costs associated with drive purchase, anything that can be used to extend the useful life of drives will yield both carbon benefits for the planet and economic benefits for the SNO. However, most drives are taken offline after 3-5 years, both because the risks of drive failure increases over time and because most drives go out of warranty.

Given the paramount importance of data durability, it is unsurprising that organizations that store data using traditional methods are unwilling to store important data on older drives that are out of warranty. However, Storj has built a system that assumes drive failure as a normal part of the data lifecycle, and protects against failure using erasure coding. As described above, a segment is divided into 80 erasure coded pieces, of which any 29 can be used to reconstitute the segment. Each of those 80 pieces is stored on different drives in different parts of the network. As a result, approximately 80-29=51 drives would have to fail simultaneously to impact durability. That is almost statistically impossible for any given segment given the diversity of geographies, operators, power supplies, etc. serving those 80 pieces.

A drive with a 5% chance of failing during a year represents an unacceptable risk for an enterprise using that drive to store a single copy of sensitive data. However, the same drive represents almost no risk to a network like Storj's where data recovery is not dependent upon any one drive. As a result, it makes economic, durability, and carbon sense to extend the life of drives well beyond their 3-5 year lifespan.

#3: Reduction of copies needed for durability

As noted in the section about data growth, most hyperscalers store a minimum of three copies of data per region following the 3-2-1 Backup Rule using replication to achieve acceptable levels of durability. That number can go much higher as the expectation for hard drive failure increases. And, because many corporate data centers are subject to greater risk of failure than hyperscale data centers, many corporate data centers need to store more than 3 copies.

Storj does not rely on replication to achieve durability and distribution, but instead uses Reed Solomon[17] erasure coding. Storj is able to achieve over 11 9s of durability with an effective expansion factor of 80/29= 2.7. That includes providing multi-region protection against data center or even geography wide risks such as fires, floods, power outages, and civil unrest.

#4: Reduction of copies needed for geographic distribution

Many organizations not only store multiple copies of data, but also store copies in multiple locations in order to avoid the risks that can affect single locations. Multiple copies are also often required if data is to be performantly accessed in multiple locations. Distributing data to a location that is far from where it is stored can take a lot of time, due to factors such as the speed of light, the number of "hops" required, and congestion in parts of the network (e.g. transoceanic cables). Thus, in order to achieve acceptable levels of geographic performance, organizations will often store multiple copies of the same data in data centers around the world, with the

[17] Reed-Solomon erasure coding definition

accompanying economic and carbon costs.

In the case of Storj, pieces of a given file are globally distributed. To download a segment, only the fastest 29 out of 80 drives need to respond. Thus, for example, a person viewing a video in Mumbai will likely get the video from 29 different drives than the person viewing that video in Memphis—who will get it from different drives than the people in Melbourne, Marrakesh, or Montevideo. Storj is able to deliver globally good and consistent performance without increasing expansion factor, thus reducing both economic and carbon costs.[18]

#5: Efficient distribution between HDD and SSD

Many use cases call for the use of both HDDs and SSDs. For example, a typical video distribution use case will have the entire library of content deployed on HDD-based origin servers, replicas of which are placed in multiple geographic locations. Closer to the edge, content that is especially “hot” will be stored in SSD-based cache or Content Delivery Networks (CDNs). However, the overall life cycle carbon costs of SSDs are nearly twice that of HDDs.[19] As a result, there are significant economic and carbon savings to be had in enabling an efficient distribution of storage between HDD and Solid State Drives (SSDs).

The Storj solution significantly reduces the need for multiple, HDD-based origin servers, because there is a globally distributed, globally perforant base storage layer. However, the speed of the Storj system also means that a significant percentage of the library content can be served performantly from the origin itself, rather than requiring SSD-based caching. While SSD-based caching is often still desirable for the most popular content, the amount of SSD-based storage can be reduced significantly, with attendant carbon and economic savings.

Note: These savings are not captured in the current carbon model calculations as this may differ by storage use case.

[18] Why cloud storage has inconsistent performance and how to fix it.

[19] The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

Methods

A model to estimate carbon savings

Comparing the carbon impact of mining rare earth elements versus powering a drive involves many good faith estimates. Like most other carbon savings calculations it is not possible to precisely measure the savings, but we believe our calculations are a conservative and robust estimate of the savings from using Storj. The numbers in the model represent good-faith estimates that can be used to estimate the CO_2 savings from using the Storj network specifically, and from similar attempts to better utilize hard drives generally.

Figure 7: Model for calculating data storage carbon output with default assumptions.

Carbon Impact of Various Storage Modalities						
	Traditional Modalities		Storj Modalities			
	Hyperscaler	Corporate DC	Storj Standard*	Storj Reused*	Storj New Nodes	Storj Blended
Drive Lifetime Carbon Analysis						
Drive Lifetime (Years)	4	4	6	3	6	**5.4**
Carbon impact from mining, manufacture, transport (kg CO_2)	20	20			20	**11.6**
Yearly Carbon impact from mining, extraction, manufacture, transport	5	5	0	0	3.3	**1.9**
Carbon from power per year of operations (kg CO_2)	15.9	15.9		15.9	15.9	**12.6**
Incremental Carbon to Write 1 TB (kgCO_2)			3.5	3.5	3.5	**3.5**
Amoritized Lifetime Carbon from Write (kg CO^2)			1.2	1.2	1.2	**1.2**
Yearly repair factor			1.18	1.2	1.2	**1.2**
Additional overhead for Storj Satellites (kg CO_2)			0.9	0.9	0.9	**0.9**

*Storj Standard: already provisioned, powered drives

*Storj Reused: Old drives brought back online

Carbon Impact of Various Storage Modalities						
	Traditional Modalities		Storj Modalities			
Yearly Cost per TB at 100% utilization (kg CO_2)	20.9	20.9	2.3	18.0	21.3	**16.6**
Utilization (%)[20]	75%	40%	85%	85%	85%	**0.9**
Power Mix Factor			74.5%	74.5%	74.5%	**0.7**
Effective Carbon Cost/TB-Year (kg CO_2/TB)	**36.2**	**67.9**	**2.0**	**15.8**	**18.7**	**14.6**
Storj Network Weighting						
% of Storj Network			0.21%	0.208%	0.582%	**100%**
Customer Specific Factors						
TB of data (TB)	1.0	1.0	1.0	1.0	1.0	**1.0**
Years Stored (Years)	3.0	3.0	3.0	3.0	3.0	**3.0**
Replication for Durability (expanded data/source data)	3.0	4.0	2.7	2.7	2.7	**2.7**
Multi Region Factor (# of regions)	2.0	2.0	1.0	1.0	1.0	**1.0**
Total at Full Capacity (kg CO_2)	**326**	**679**	**15**	121	144	**112**
Carbon Savings						
Carbon Savings from Storj Std vs. Alternative (kg CO_2)	**311**	**664**				
Carbon Savings from Storj Bended vs. Alternative (kg CO_2)	**214**	**552**				
Carbon Savings from Storj Std vs. Alternative (%)	**95.3%**	**97.7%**				
Carbon Savings from Storj Bended vs. Alternative (%)	**65.6%**	**83.5%**				

*Storj Standard: already provisioned, powered drives

*Storj Reused: Old drives brought back online

[20] Estimates of data center and storage utilization vary widely. A few sources include: C&C Technology Group, Natural Resources Defense Council (NRDC) and IBM. To keep our calculations conservative, we used the higher range of estimates for utilization, which makes the comparative results for Storj less favorable than they might be otherwise.

Results

Using the assumptions noted in Appendix A, the model suggests that using Storj in its pure form to store a TB of storage for 3 years generates 12 kg of CO_2. Storing that data with a hyperscaler generates 251 kg of CO_2. Using a corporate data center generates 523 kg of CO_2.

That means that utilizing Storj for data storage results in a reduction of 239 kg of CO_2 per TB compared with hyperscaler storage and a reduction of 511 kg of CO_2 compared with a corporate data center. This is a 95% to 98% reduction respectively.

To put that into perspective, a car produces about 0.24 kg/km of travel.[21] So, for a single TB of effective storage, the equivalent of between 1000 km-2000 km of emissions from car travel can be eliminated.

Of course, these are good faith estimates only, but do suggest the power of this model. If just a fraction of the ZB of data expected to be produced in the near future and the ZB of capacity in already deployed and powered drives around the world could be switched to Storj, the impact on the environment would be significant.

Figure 8: Summary of the results from the model for the carbon impact of data storage.

SUMMARY TABLE				
Carbon Impact of Various Storage Modalities				
	Traditional Modalities		**Storj Modalities**	
	Hyperscaler	**Corporate DC**	**Storj Standard***	**Storj Blended**
Total at Full Capacity (kg CO_2)	362	679	15	112
Carbon Savings from Storj Std vs. Alternative (kg CO_2)	311	664		
Carbon Savings from Storj Blended vs. Alternative (kg CO_2)	214	552		
Carbon Savings from Storj Std vs. Alternative (%)	95%	98%		
Carbon Savings from Storj Blended vs. Alternative (%)	66%	83%		

[21] Carbon Dioxide, Part 2: Walk, Drive a Car, or Ride a Bike? | GLOBE Scientists' Blog.

Discussion

66-83% carbon savings can be achieved

The global datasphere is large and growing rapidly, and the storage and distribution of that data already represents a significant and growing contributor to global carbon.

Both the carbon impact from the manufacture and transport of storage devices and their ongoing operation represent significant sources of carbon. The fact that most drives are severely underutilized and the fact that it takes almost no additional electricity or cooling to run a drive at near full capacity versus partial capacity suggests that one of the easiest ways to achieve significant carbon savings is by enabling already powered and deployed drives to be more fully utilized. Additional significant savings can be achieved by extending the effective life of already manufactured drives, minimizing the number of copies of data that need to be stored to achieve equivalent data durability levels, minimizing the number of copies of data that need to be stored to achieve equivalent global geographic distribution and performance levels, and enabling a carbon efficient distribution of storage between HDD and SDD.

Storj is a system that is designed to provide enterprise grade, S3 compatible cloud storage by leveraging underutilized capacity from independently operated drives around the world. While acknowledging that accurately comparing carbon contributions from disparate and heterogeneous systems is difficult, the model presented is tunable to individual users' parameters. The assumptions used in the model result in a significant carbon savings of 66-83% relative to hyperscalers and even greater savings relative to less efficient data centers.

For green solutions to be effective, however, they must deliver real economic value. And, economic incentives must be aligned with green outcomes. The Storj system is designed such that storage node operators are incented to more fully utilize already powered and deployed drives and extend their useful lives. And consumers of data storage are incented to use the system because of favorable economics combined with favorable durability, security, and performance characteristics. All of these benefits, both green and economic, are inherent to the distributed nature of the service.

Appendix A

Assumptions

Carbon impact of running drives

The model starts with an assumption of the lifecycle costs of manufacturing drives. Tannu and Nair (2022) analyzed the LCA (Life Cycle Analysis) of 24 HDDs using manufacturer data. Their estimates of 20 kg CO_2 per effective TB were used in this model. This represents an average over a number of different drive sizes and configurations.

Tannu and Nair's (2022) estimate of power required per year per TB for HDDs of 36.8 kWh is also applied and multiplied by their recommended conversion factor of 0.7 kg/kWh to get an annual carbon impact from powering drives of 15.9 kg CO_2.

The marginal impact of writing a TB of data uses Williams' (2017) estimate of 0.005 watt hours to write a GB of data and the aforementioned 0.7 kg CO_2/kWh to arrive at 3.5 kg CO_2/TB of data written.[22]

write (watt-hours/GB)	0.005
kg CO_2/kWh	0.7
kg CO_2/TB written	3.5

For drives in the Storj network that are already being spun and powered, the incremental carbon from writing is really the only incremental carbon associated with filling up an already running drive.

Storj does need to maintain metadata servers (called satellites) in order to appropriately manage the diverse pieces on multiple nodes, pay SNOs, and coordinate retrieval. The total current carbon footprint of those services (using Storj cloud providers' tools) and the total unexpanded data currently in the network resulted in an additional overhead factor of 0.2 kg CO2/year per unexpanded TB of data. Note that Storj's cloud provider does claim to offset carbon, but to be conservative,the unoffset numbers are used for these purposes.

[22] Stanford: Environmental impact of writing data to the cloud

Storj Meta Data Overhead	
Storj Google Carbon Footprint-Feb (kg CO_2)	3600
Annualized	43200
Total current Expanded Storage of Network	18933
Carbon Overhead TB/YR	0.19

This results in an effective carbon cost for storing a single copy of 1 TB of data Next, an assumption was made for an average lifetime for the drives. Since most enterprise drives have a lifetime of between 3 and 5 years, an average of 4 years for both hyperscaler and enterprise drives was used in the model. For reasons stated above, a slightly longer effective lifetime of 6 years was used for drives in a Storj-like network.

Considering utilization, as noted above, corporate data centers are woefully underutilized, but utilization a factor of 40% was used, rather than the 12-18% cited earlier, in order to be conservative. For hyperscalers, a 75% utilization factor was assumed. An 85% utilization factor was used for Storj.

This results in an effective carbon cost of storing a single copy of 1 TB of data for a single year of 27.9 kg CO_2 for a hyperscaler, 52.9 kg CO_2 for a corporate data center, and <2 kg CO_2 for Storj.

User centric factors

To help the model extend to a particular user's needs, users are able to specify:

1. The total amount of data to be stored.
2. The number of years to be stored.
3. The expansion factor used to ensure durability.
4. The expansion factor used for multi-region performance.

The model also allows users to set a replication/expansion factor for durability. The default option for AWS, Google, and Microsoft is to store three copies of data. Generally, the hyperscalers offer both replication on three different drives in a single data center and, for slightly more money, three copies spread between different data centers within a geographic region. All three claim to have durability at or above 11 9's with this configuration. For storage within a corporate data center, it is unlikely that the same levels of durability can be achieved with an expansion factor of three, given the inability of most small data centers to invest in the same level of monitoring, management, fire suppression, and infrastructure as the hyperscalers. To be accurate, a person choosing to store data in their own data center would have to make an assessment of the durability achievable given the MTTF characteristics of their drives, environmental factors, etc. For the purpose of the default model, a corporate expansion factor of 4 and a hyperscaler expansion factor of 3 is assumed.

Storj is able to provide over 11 9's of durability, with an expansion factor of 2.7. The effective expansion rate is actually slightly less (2.5) given that repair is only triggered when the number of healthy pieces in a segment falls below the repair threshold. Nevertheless, the more conservative number is used for the purpose of this analysis.

In addition to storing multiple copies for durability purposes, some may choose to store copies in different regions/geographies in order to provide good performance for consumers of data in different locations. For the purposes of the default model, a geographic replication factor of 2 (i.e. 2 different geographic locations) is used. This will differ widely by use case and it is reasonable to assume that some users may reduce the number of copies stored in each region, relying on multi-region copies to provide both performance and durability.

Storj is able to provide multi-region performance without needing to increase the expansion factor.

All of these user factors are multiplied by each other and the kg/TB-year factor above to get a total carbon outcome.

Actual composition of the Storj network

The Storj network is primarily composed of SNOs who are using Storj to increase utilization of nodes that are already provisioned and being powered. Based on surveys, this represents 69% of the current network–a number that we expect to increase over time, as it is the most profitable configuration for SNOs.

However, some (roughly 16%) have brought previously inactive nodes back online for a few years. While these nodes don't incur net new carbon related to manufacturing, the model assumes that the full carbon cost of power of running the drives should be assigned. It is also assumed that these drives will have a lower effective lifetime, as most are already a few years old.

Finally, some percentage of SNOs (approximately 15%) have created new nodes specifically for the purpose of earning rewards. While this is not actively encouraged, and is the least profitable configuration, nothing prohibits this configuration. For this part of the network, the full life cycle carbon costs have been assigned to the drive.

Weights of 69%, 16%, and 15% were assigned to these three configurations, and reflect a blended carbon impact of the Storj network as a whole. While this increases the net carbon contribution of the current Storj network, it still compares quite favorably. It is worth noting that the percentage of nodes that are using already deployed and powered nodes is expected to increase as economic incentives are adjusted in the future to make this style of node operation the most profitable.

Appendix B

Model refinement

1. **Data reads** - The model does not consider the carbon impact of data reads or transmission, partly because it is difficult to find good source data, partly because different workloads vary considerably in their read intensity, and partly because the carbon impact of reads from different storage are unlikely to differ materially.
2. **Reduced reliance on SSDs** - A related topic is the extent to which different storage configurations can reduce the reliance on Solid State Drive (SSD) storage for high performance reads. As Tannu and Nair (2022) note, while SSDs do require less power to operate than HDDs, the manufacture of SSD is significantly (4x) more carbon intensive, in addition to being far more expensive on a per-TB basis. Thus, it seems likely that the Storj solution, which can performantly serve a long tail of content directly from origin servers, without the need for cache-based CDN, might yield both carbon and economic savings.
3. **Fuel mix/energy sources** - It is worth noting that the mx of power used (e.g. fossil fuel vs. renewables) has a huge impact on the carbon intensity of the power stage. Thus, results could differ considerably in different areas of the world. In areas of the world with an abundance of clean power, the relative impact of manufacturing/mining/distribution is likely to be even greater.
4. **Enterprise vs. prosumer drives** - The Storj system currently comprises a mix of drives operated in data centers and prosumer drives. Life cycle analysis suggests that the characteristics of these types of systems differ on several important dimensions. Enterprise drives tend to be higher capacity, involve more intense manufacturing, and are run in active mode a far greater percentage of the time relative to prosumer drives. It is unclear whether the net impact of these factors would cause the better utilization of already deployed drives in data centers to cause greater or less carbon savings than the better utilization of prosumer drives.
5. **Cooling impact** - One of the largest sources of power usage in data centers is cooling. While the percentage of data center power used by cooling is noted in the paper, cooling does not factor into the current model. It is reasonable to assume that the additional heat produced by a drive is roughly proportional to

the amount of power it consumes. Therefore, if cooling and other data center costs were included, they would likely serve to further emphasize the carbon savings from better utilizing existing, powered drives.

6. **Other configurations** - The Storj technology can combine underutilized storage from independent operators around the global into a unified object storage. Its current instantiation is a market based model that relies on zero knowledge principles and a broad and diverse set of storage node operators. Storj as a product is just one possible usage of the Storj technology. Other products could be built (e.g private or restricted networks) utilizing storage space from more trustworthy and less heterogenous sources. These models could allow further optimizations, including Reed Solomon ratios with an even lower expansion factor, improved performance to further reduce the impact of SSD, etc. These models should be explored, as they could be more optimal and thus even greener than Storj.

Appendix C

References

1. Worldwide IDC Global DataSphere Forecast 2022-2026

2. US Chamber 3-2-1 Backup Rule

3. Replication calculations for durability - Pages 21-22

4. C&C Tech Group - Understanding Data Center Energy Consumption

5. IEA Data Centres and Data Transmission Network Report

6. Shehabi et. al. (2016). United States Data Center Energy Usage Report

7. Seagate life cycle analysis of the Momentus drive and impact on climate change

8. Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

9. Mostafa Sabbaghi, et al, The Global Flow of Hard Disk Drives: Quantifying the Concept of Value Leakage in E-waste Recovery Systems.

10. Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

11. Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

12. The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

13. Hongyue Jin, et al, Life cycle assessment of emerging technologies on value recovery from hard disk drives

14. Are Your Data Centers Keeping You From Sustainability?, IBM

15. The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

16. Stanford Business: Vinod Khosla - Green Tech Must Make Economic Sense

17. Reed-Solomon erasure coding definition

18. Why cloud storage has inconsistent performance and how to fix it.

19. The Dirty Secret of SSDs: Embodied Carbon, Swamit Tannu University of Wisconsin-Madison and Prashant J Nair University of British Columbia.

20. Estimates of data center and storage utilization vary widely. A few sources include: C&C Technology Group, Natural Resources Defense Council (NRDC) and IBM. To keep our calculations conservative, we used the higher range of estimates for utilization, which makes the comparative results for Storj less favorable than they might be otherwise.

21. Carbon Dioxide, Part 2: Walk, Drive a Car, or Ride a Bike? | GLOBE Scientists' Blog

22. Stanford: Environmental impact of writing data to the cloud

STORJ

Storj
1450 W. Peachtree St. NW #200
PMB 75268
Atlanta, GA 30309-2955
info@storj.io