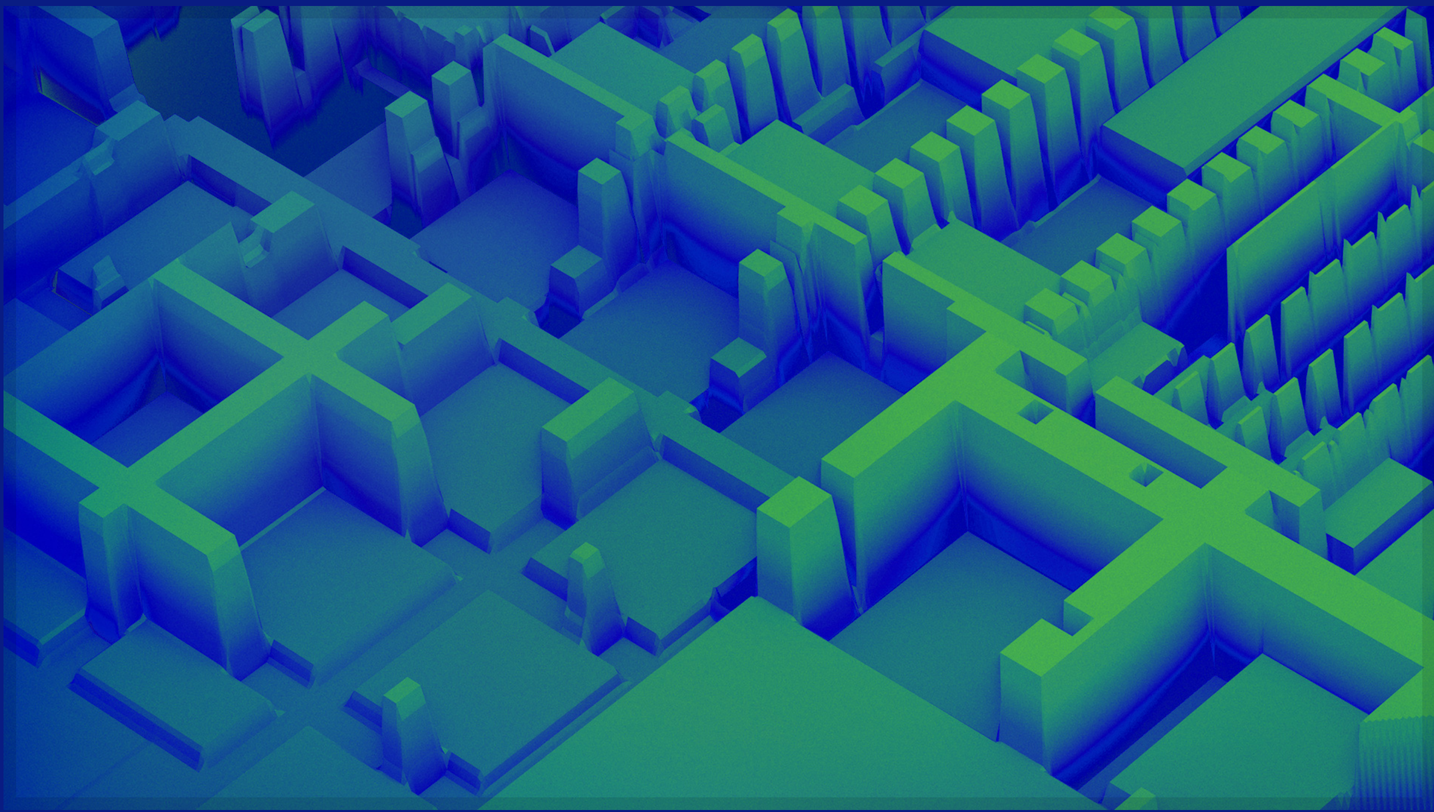
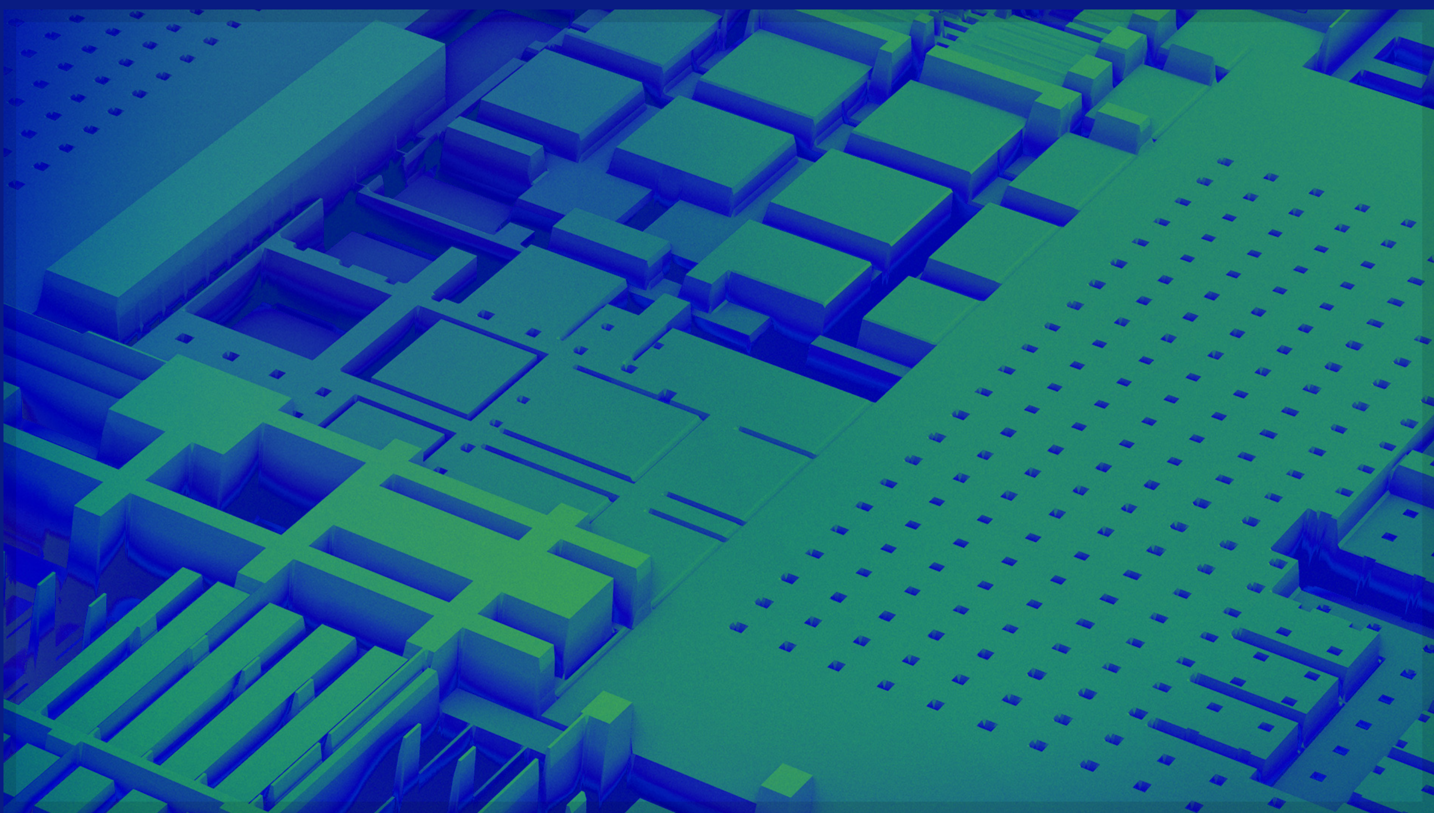
THE BLOCK · Research

COMMISSIONED BY W3BCLOUD™



DECENTRALIZED STORAGE: A PRIMER



Commissioned by W3BCLOUD™

W3BCLOUD™

W3BCLOUD™ (pronounced Web3Cloud) is a leading infrastructure provider powering Web3. We believe everything that can be decentralized will be decentralized, and we provide robust infrastructure required to scale these decentralized protocols and their applications. In short, we are the AWS of Web 3, providing enterprise grade infrastructure that is optimized for the next generation of the Internet.

Researched by The Block Research

THE BLOCK · Research

The Block is an information services company founded in 2018. Its research arm, The Block Research, analyzes an array of industries including digital assets, fintech, and financial services.

Contact

Email: research@theblockcrypto.com
Twitter: @theblockres

Author

Hiroki Kotabe, Research Analyst
Twitter: @blockhiro

Table of Contents

Section 1: Inception & Evolution of Decentralized Storage

Digital storage has come a long way. Before the 1900s, we were using punched cards to represent digital information – stiff paper with information represented by the presence or absence of holes. In 1928, IBM introduced a new version of the punched card with a rectangular hole and 80 columns.

It took a competition between two of IBM's top research teams to come up with this new digital storage technology. It propelled them to the forefront of data processing – for almost four decades, this was the major way to store, sort, and report data.

It wasn't until 1956 that IBM released the first hard drive. It was bigger than a refrigerator and weighed more than a ton. And it could store roughly 5 megabytes (MB) of data, the equivalent of about one MP3 music file. But, in the 1950s, it was a massive achievement.

The floppy disk arrived in 1971, again from the labs of IBM. The first floppy disk was huge at about 8" and stored just 100 kilobytes (KB), but about a decade later Sony introduced the modern 3.5" floppy disk which could store 400 KB to 1.44 MB of data.

It wasn't until the late 90s and early 2000s when we started seeing modern forms of portable storage like DVDs and SD Memory, capable of holding several or several dozens of GB, respectively.

But it was in 2006 that the world of digital storage evolved from local portable storage to global online storage, with the rise of cloud storage services like Amazon S3.

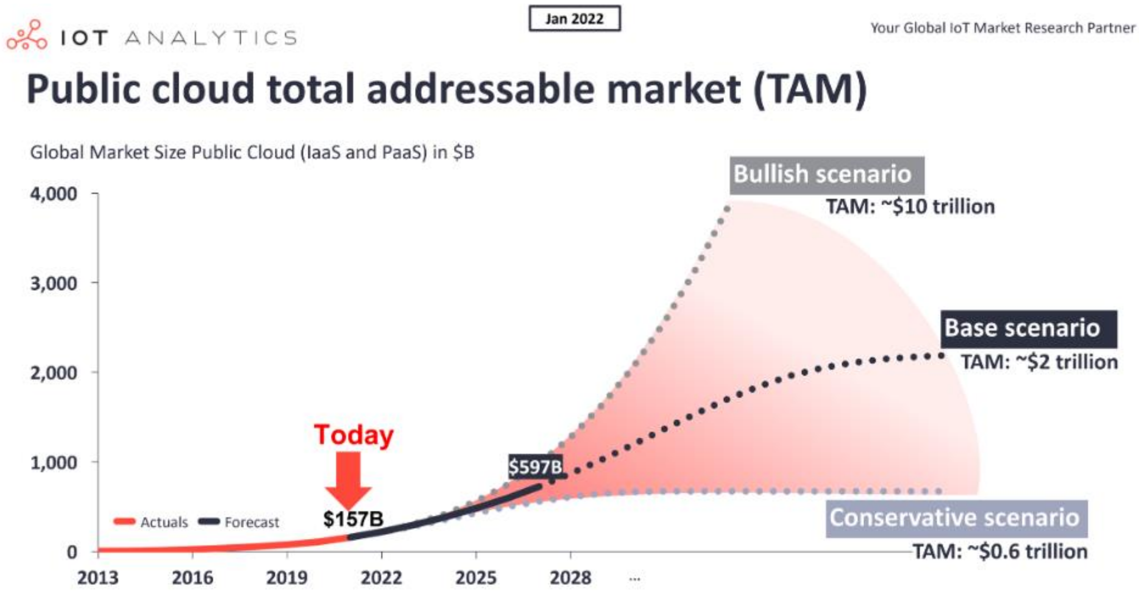
Despite the increasing pace at which revolutionary developments have been achieved in storage, there hasn't been a major revolution since the advent of cloud storage services. But, it appears that the next revolution might be in peer-to-peer (P2P) cooperative storage clouds – aka decentralized storage networks.

In the early 2010s, computer scientists like Juan Benet started building the foundations of decentralized storage networks. He founded the development firm Protocol Labs in 2014 and in 2017 raised around $205 million in one of the largest token offerings at the time. Protocol Labs continues to develop the most successful decentralized storage protocols – the InterPlanetary File System (IPFS) and Filecoin.

Filecoin and other decentralized storage protocols like Sia and Storj arose as part of a movement started by Bitcoin born out of the turmoil of the 2008 Great Recession. People's distrust of banks and their role in the financial system grew as they watched the house of cards built out of hubris and wretched excess fall.

Similar to how Bitcoin is disrupting traditional banking systems by operating a currency with P2P software and cryptography instead of trusted intermediaries (e.g., banks, governments), decentralized storage protocols like Filecoin are disrupting traditional cloud storage by operating a cooperative storage cloud with P2P software and cryptography instead of trusted intermediaries (e.g., AWS).

Also like Bitcoin, the total addressable market (TAM) is huge. One in-depth [study](#) makes a case for a $2 trillion TAM for public cloud computing, with an optimistic scenario reaching dizzying heights of ~$10 trillion and a conservative scenario at ~$0.6 trillion. These estimates take into account a market size of $157 billion in 2021, with practitioners estimating about 15-25% of workloads in the public cloud at that time and 50-75% of workloads in the public cloud in a future end state.



Source: IoT Analytics

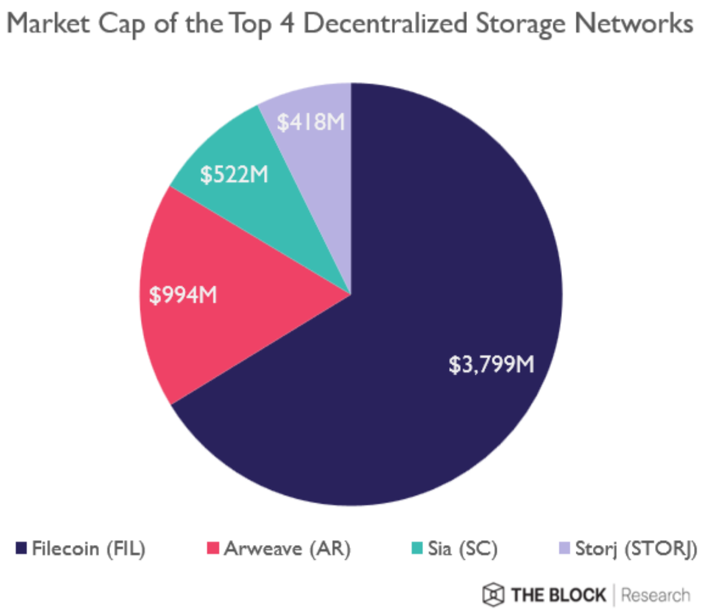
Realizing the vision to store the planet's most important data in decentralized storage networks is not only a mission for the public good, but also for capturing a market worth potentially hundreds of billions of dollars. We see that there is a massive opportunity for web3 to capture significant value by deploying apps in this space. To motivate this, it's important for people to understand the transformation in computing that is happening behind the scenes as we shift from web2 to web3 storage. The remainder of this report will focus on explaining this shift.

Section 2: State of the Market

Blockchain-enabled decentralized storage has gained interest and adoption at an explosive rate since early 2021, coinciding with the rise of the mainstream NFT market and surging interest in web3. Developers are rapidly becoming aware of the new way to store and process data that doesn't rely on centralized services like AWS.

The new generation of blockchain-enabled cooperative storage clouds not only fit the decentralization ethos of web3, but also offer potential technological advantages over their centralized counterparts. It's still the early days, but several of these protocols are off to a promising start.

Filecoin is the most prominent entity in this emerging market in terms of storage usage and capacity. Filecoin has been described as a "P2P version of AWS," perhaps rightly so as it has accumulated more than eightfold the storage usage of its second-leading competitor Storj. Furthermore, it vastly exceeds its competitors in terms of market cap, as shown in the table below.



Source: CoinMarketCap, 2022/04/20

So, to understand how decentralized storage networks are useful and what is driving this burgeoning market, we use Filecoin as a case study. We cover relevant aspects of its closest competitors Sia and Storj, and include a separate section for Arweave as it's the leading alternative protocol addressing the need for permanent storage instead of contract-based temporary storage.[1]

We focus on the technologies powering these decentralized storage protocols as they shed light on how they can achieve the two ultimate goals of any data storage service:

- **Data availability** – means that stored data is quick and reliable to access and use
- **Data integrity** – means that data accuracy and consistency are maintained over its entire lifetime

To understand how decentralized storage networks can achieve these goals, it's important to understand storage and retrieval deals and the technologies powering decentralized storage networks. Let's begin there.

[1] The top six protocols tagged with "storage" by CoinMarketCap also include BitTorrent (BTT) and Holochain (HOLO), but we do not discuss them here for the following reasons: BitTorrent is a torrent client used as a P2P network for sharing individual files – it does not follow the "single swarm" design of decentralized storage whereby all files are hosted by a single network. It is more relevant to a discussion about torrents than decentralized storage in web3. Holochain does include a distributed storage mechanism, but as part of a comprehensive project trying to create a general framework for distributed P2P computing. Also, it is not a blockchain-based protocol.

Section 3: How Decentralized Storage Works

A defining feature of decentralize storage within a technological framework of web3 is the use of a *permissionless* network of storage nodes, meaning that anyone can add storage capacity to these networks and earn revenue.

The value add of a permissionless network is it creates an *open market* of storage providers. While some protocols like Filecoin, Sia, and Storj use this open-market function more explicitly by actively incentivizing redundancy, others like Arweave use it more implicitly by assuming that open-market competition will bring about storage redundancy.

Underpinning all these protocols is the web3 ethos of openness and decentralization, which enable several properties that cloud consumers care about to varying degrees depending on their use case:

1. **Censorship resistance** – No individual entity can take down content, the open-market approach implies an architecture where data is not locked to a single provider.

2. **Resilience** – Data integrity is guaranteed by storage redundancy produced by the open-market approach. Redundancy is important to secure data given the more hostile conditions than a traditional cloud storage environment.

3. **Cost** – Ideally, it will be lower cost than comparable centralized cloud services due to pricing efficiency driven by open-market competition.

One of the main obstacles facing progressive decentralization in web3 is about how to store large amounts of data in a decentralized way, given the prohibitive costs of storing data directly on popular blockchains like Bitcoin and Ethereum, which are not designed for storing large files (e.g., NFT metadata and hypermedia). For example, it would cost about ~$20M to store 1 GB of data on the Ethereum blockchain.

Bitcoin and Ethereum intentionally target a block size and a state growth rate to enable thousands of validator nodes. These targets require a level of computational and storage redundancy that make them unfit for large cloud storage – fault-tolerant decentralized consensus across requires data replication, processing, and storage across all full nodes.

In contrast, storage chains like Filecoin separate the consensus aspect of their blockchain from the storage of files. Cryptographic securities and economic structures are used to align incentives and write the storage state to a shared consensus that can be verified by users. Effectively, storage chains like Filecoin enable a consensus structure for each storage network and cryptographic and economic designs aim for service guarantees and data availability and integrity.

As previously mentioned, the current leaders in the decentralized storage market are Filecoin, Sia, Storj, and Arweave. Filecoin, Sia, and Storj are *contract-based* decentralized storage protocols in that storage buyers and sellers negotiate temporary storage deals in open markets. In contrast, Arweave is an *incentive-based* and *permanent* decentralized storage protocol. Arweave attempts to address the limitations of storing data on blockchains by utilizing a "blockweave," a blockchain derivative that is designed specifically for handling large amounts of data storage. We cover Arweave in more detail in Section 6.

Contract-based decentralized storage networks store data off chain. Their protocols generally have a similar structure where:

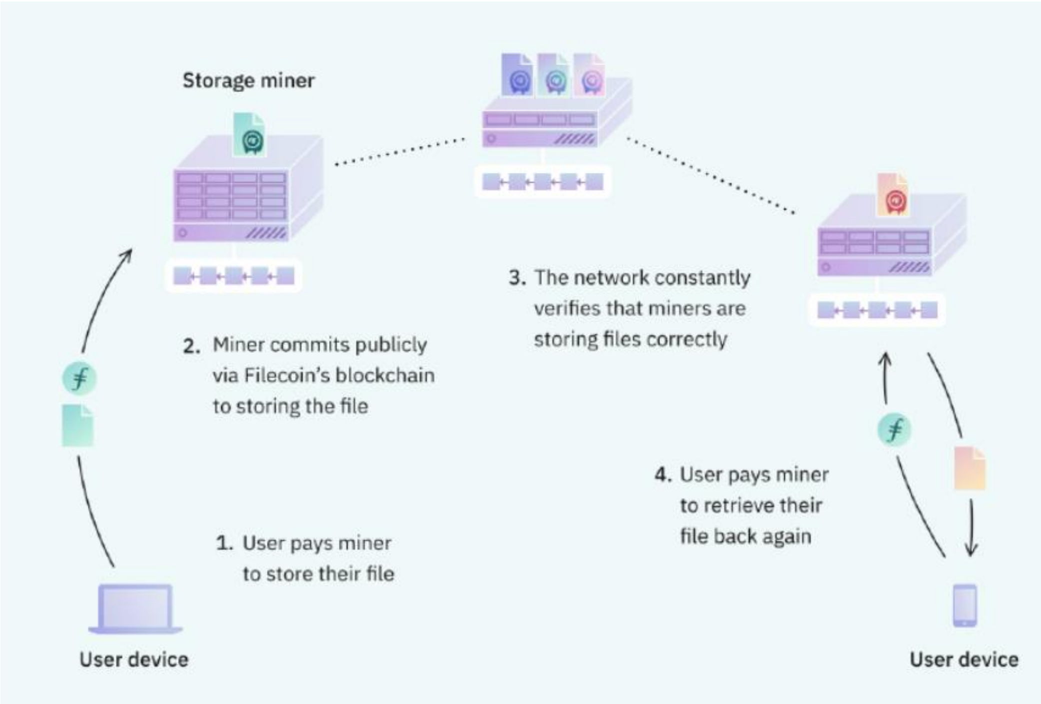
1. Storage buyers (demand) pay storage providers (supply) for storing and retrieving files off-chain.

2. To maximize availability and fault tolerance, these files are usually split into chunks stored across multiple storage providers for redundancy or they utilize error-correction technology. Some protocols encrypt the chunks themselves (content encryption) while others encrypt only when sending data (transport encryption).

3. Since storage providers are not trusted entities, they are required to collateralize (e.g., via staking) or be vetted in some way for the opportunity to work in the network.

4. A system of rewards (e.g., mining rewards, gaining network power) or punishments (e.g., withholding rewards, slashing, losing network power) is used to try to ensure service guarantees.

For a better understanding, let's break down how storage and retrieval deals work within Filecoin's P2P open-market system depicted below.



Source: Filecoin

Storage Deals

Storage deals are agreements made between clients and miners about the cost and duration of data storage.

The lifecycle of a storage deal follows five phases:

1. **Discovery** – Clients identify miners to get pricing information (e.g., price per GB per minute).

2. **Negotiation** – Client and miner come to an agreement about the terms of the deal, including fees, storage duration, starting time, and so forth.

3. **Data transfer** – Data is then transferred from client to miner.

4. **Publishing** – The deal and its terms are published on chain, making the miner serving the client publicly accountable for servicing the deal.

5. **Handoff** – Miners then pack client data into a sector of Filecoin's storage mining sybsystem, prove once on chain that they have stored a full and unique copy of the data (a computation-heavy process known as "sealing"), and then later prove continuously on chain that they are upholding the storage deal throughout its lifetime.

Unlike centralized storage services, the storage cost is determined by free-market dynamics in an open market instead of a fixed pricing structure. To incentivize storage providers to participate in storage deals, the Filecoin network rewards storage providers with the network's native token (FIL). Storage providers earn block rewards for committing storage for future deals on top of payments for storage and retrieval.

Retrieval Deals

In Filecoin retrieval deals, clients pay miners to fetch their data via off-chain payment channels. Payment channels are a kind of state channel where parties can lock their funds into an on-chain contract, transact directly with each other off-chain, and settle their balances back on-chain – thereby deriving security from the underlying blockchain.

The use of payment channels helps clients retrieve their data fast while keeping costs low. Unlike on-chain payments, payments via payment channels don't require validation by or communication to the broader network. Payment channels are the technology powering Bitcoin's Lightning Network.

Data transmission is metered – the storage buyer pays the storage provider incrementally as the data is being transferred. Opening channels, closing channels, and redeeming retrieval vouchers are the only parts of the process that interact with the Filecoin blockchain.

The overall process is as follows:

1. **Find data** – The client identifies miners storing the data they need and request a retrieval quote from them, including price per byte, price for unsealing if needed, and payment interval.

2. **Set up payment channel** – The client sets up a payment channel between them and the miner if it's not already in place.

3. **Transfer data and pay** – Miner sends the data to the client until payment is required based on the payment interval agreed in the Discovery phase – payments are requested after a data transfer threshold is met, and then the rest of the data transfer is completed after payment.

Addressing and Sharing

Storage and retrieval deals typically take place on top of a P2P addressing and sharing layer. Addressing refers to a mapping of content identifiers to the content the identifiers are derived from. It is an essential mechanism for sending and receiving the correct content.

In Filecoin's case, that addressing and sharing layer is a separate complementary protocol called the InterPlanetary File System (IPFS). Simply put, IPFS addresses and moves content while Filecoin creates incentives to persist data.

By separating these protocols, IPFS can technically work without Filecoin, with IPFS used more for data availability while Filecoin is used more for guaranteeing data persistence. For example, Filecoin could use a different P2P addressing and sharing layer and IPFS could work without an incentive layer through self-organized or altruistic forms of data persistence. For example, the IPFS cluster which powers large storage services like NFT.Storage and Web3.Storage operates nodes independently of Filecoin. In contrast, Sia and Storj draw less distinction between their incentive system and their addressing and sharing system.

Another way to think of IPFS is that it is like a P2P version of HTTP in that both are used for addressing and moving content. See the table below for a comparison between these two technologies.

HTTP vs. IPFS

Critera	HTTP	IPFS
Centralization	Centralized client-server approach	Decentralized peer-to-peer approach
File Retrieval	Location-based – data is requested using the address of the physical location where the data is stored	Content-based – data is requested using the cryptographic hash of that data; the network only needs the hash of a file to locate the nodes storing that file
Availability	Depends on the availability of the servers where the data is stored; data is inaccessible if the server is down or any link to that server is broken	Data is copied to multiple server nodes; with high redundancy, this means little to no downtime
Bandwidth	Bandwidth is limited by the bandwidth of the centralized server that receives requests from multiple clients	Bandwidth is limited by whether there is a close or low-latency peer storing the data
Cost	Clients pays the hosting server; costs are increased due to location and infrastructure requirements	Clients pay to store data at specific levels of redundancy or availability
Adoption	HTTP is well-established and inbuilt on almost all computers	IPFS is relatively new and less popular and can only be accessed by installing an HTTP-to-IPFS gateway or node

Sources: The Block Research

IPFS vs. Torrents

IPFS can also be likened to torrents in that they are both decentralized and utilize content addressing. One of the major differences is that IPFS uses a single massive P2P swarm, where anybody can move content to anybody, helping to reduce data-retrieval latency. On the other hand, torrents use smaller P2P swarms for each individual torrent file. For example, with torrents, two files with only slightly different data won't share peers as IPFS does. They will each possess their own P2P swarms. Then, downloading IPFS files can be more efficient and significantly faster than torrent files due to the larger network of peers hosting the requested data blocks.

So far, we have defined decentralized storage in the context of web3, discussed the essential workings of storage and retrieval deals in contract-based decentralized storage, and distinguished the addressing and sharing layer of decentralized storage from incentive-based persistence networks and torrents. Next, we will go into technical details that reveal how the tech powering decentralized storage supports data availability and integrity via data security, redundancy, and efficiency mechanisms.

By looking at these technical aspects of decentralized storage, we can start to see the unique offerings of this alternative model of cloud storage.

Section 4: The Tech Powering Decentralized Storage

Decentralized storage is a model of networked online storage where data is stored on multiple computers (nodes) hosted by participants cooperating to create an emergent cloud. The capacity for flexible contribution allows some nodes to contribute less storage and others to contribute more. Incentives can be used to encourage greater storage contribution and network participation. Ultimately, the model can work if the total storage supplied in aggregate is greater than or equal to the storage demanded. The market viability depends on whether the cost of decentralized storage is less than or equal to comparable centralized services.

Unlike centralized storage clouds, building a decentralized storage network only requires marginal investment costs determined by an open market of participants rather than a single entity incurring the costs for all dedicated storage hardware. In decentralized storage clouds, the costs for dedicated hardware are potentially much less, and the base over which those costs are borne is substantially more distributed. In theory, a network of rational actors should be able to coordinate to build a substantially larger and cheaper cloud than what is offered by their centralized counterparts.

Each node in a decentralized storage model runs specialized software which communicates with some orchestration software. The orchestration allows nodes to consume and contribute storage space in the decentralized cloud. It may or may not be centralized.

The decentralized storage model offers a variety of benefits that may make it more appealing than centralized cloud services in some use cases. The order of the following subsections do not imply importance as different use cases benefit in different ways.

On Data Security

Files hosted in decentralized storage networks are typically split into chunks and encrypted before leaving the local machine. In IPFS' case, content encryption is optional while transport encryption is enforced to prevent third parties from viewing data in movement. But users can compress and encrypt files before copying them repeatedly to the cloud to add security and reduce storage requirements.

Those encrypted or unencrypted chunks are then distributed in a (pseudo)random manner using algorithms that combine load balancing and geo-distribution to reduce latency. Typically, some form of hashing is used for addressing content, allowing data to be independently verifiable and immutable.

On Data Redundancy

To achieve data availability and integrity across potentially unreliable nodes hosted by strangers over a wide area network, the source node will add some level of redundancy to each data chunk. Then, the system can recreate the entire block even if some nodes go offline (e.g., due to loss of network connectivity, hardware failure, the computer being switched off).

Erasure coding is a sophisticated form of redundancy that is both storage- and bandwidth-efficient due to data compression. Erasure coding takes a file or object and splits it into multiple fragments while creating additional fragments called "parity blocks" that are used for data recovery as they are algorithmically derived from the original data fragments.

A less computationally intensive but more storage-expensive form of redundancy is simple duplication – this is the method currently used by Filecoin and other major decentralized storage networks. However, there are advantages to the simple approach over erasure coding:

1. The network can keep data intact enabling use cases like compute over data.

2. It gives users a choice to implement erasure coding client side rather than requiring them to do so.

3. By keeping data intact, data redundancy and renewals can be optimized based on actual data availability rather than probability bounds used in erasure coding.

On Trustless Service Guarantees

One of the key benefits of blockchain-enabled decentralized storage is the ability to secure data without requiring users to trust in a third party handling that data. For example, Filecoin achieves this through two related functions: economic penalties and zero-knowledge proofs.

Basically, if a storage provider fails to prove that they are storing the data they have contractually agreed on storing, they will be penalized by having their stake slashed.

Storage providers must complete two proofs to prove they've met storage requirements: Proof of Replication (PoRep) and Proof of Spacetime (PoSt). PoRep is used in sealing the data, where a storage provider proves that they are storing a unique and full copy of the data via a series of zk-SNARK setups which together prove that the process was done correctly. PoRep takes place just once during the initial deal between the client and storage provider when the data is first stored by the miner.

On the other hand, PoSt is used daily to prove that the storage provider is continuing to store the original data over time without manipulation or corruption. In PoSt, zk-SNARKs are used to verify that random pieces of data from the full copy are still stored and available for retrieval throughout the lifetime of the storage contract. If the storage provider fails PoSt, they will lose some or all of their collateral.

The outcome is data availability and integrity via cryptographic proofs of storage instead of relying on trusted third parties like AWS to serve the data. So far, it appears to be working quite well, with Filecoin reporting 100% uptime through 2021 while AWS faced at least four major outages during the same period.

Content Addressing

Perhaps one of the biggest benefits of decentralized storage networks like IPFS over HTTP is the use of content addressing. While HTTP takes URLs to identify content by the specific physical location of the server hosting that content, IPFS hashes the content to create a content identifier (CID) – a cryptographic compression of that content into a string of characters. For example, in May 2020, the Ethereum.org website went live on IPFS with the CID:

ipfs://bafybeidoodypolrlzufnng5swfpotyhu7cdjh5vs6sc6jbuz47lprw6wfi

By default, IPFS uses the SHA-256 hashing algorithm, though it can support many other hashing algorithms. The number of characters in a CID depends on the specific hashing algorithm rather than the size of the content itself – with SHA-256, the result is 256 bits, or 32 bytes, regardless of how large the content is.

The content ID is long and not memorable, but this issue is easily addressed by combining IPFS with a name lookup system like Ethereum Name Service (ENS), where the content address above can be linked to a memorable name like Ethereum.eth.

Content addressing offers a variety of benefits over location addressing:

- **Decentralization** – Content addressed by the hash of itself instead of by a specific geolocation can be stored anywhere. When retrieving content, the network only needs the hash of that content to locate the nodes serving that content.

- **Direct access** – A communication network like the internet should allow users to directly access the data (content) they need instead of referencing intermediary physical locations (hosts) where that data can be retrieved.

- **Lightening network loads** – Caching copies of content decentrally around a network allows retrieval from the closest or lowest-latency servers instead of requiring retrieval from a topologically suboptimal geolocated server.

Key-Value Databases vs. Relational Database

Content addressing and hashing results in a massive key-value database. A key-value database is a model for storing, retrieving, and managing associative arrays – an abstract data type that keeps a collection of key-value pairs where each possible key appears at most once in the collection. In contrast to relational databases (e.g., MySQL), key-value databases (e.g., IPFS, DynamoDB) don't use tables, rows, and columns. They also don't use SQL, making them part of the "NoSQL" database family.

In IPFS, the hash is the key, and the content is a stored value of that key. The collection of all these key-value pairs creates a distributed hash table (DHT) containing a collection of records that, in turn, contain many different fields within them, each containing data.

This kind of object-oriented data structure lends itself to demanding and high-performance environments, like in handling the workloads of web-scale applications, including social networks, gaming, media sharing, and Internet of Things (IoT). Relational databases that require a well-defined schema with data normalized into tables, rows, and columns are less suitable for such environments as they create multiple intermediary steps to data access.

For example, a popular online game with thousands of concurrent users generating many terabytes of new data every day can quickly outstrip the resources of a relational database. Such web-based applications are better served by key-value databases that can handle many thousands of reads and writes per second.

To look at the differences in more detail, see the table below:

Relational Databases vs. Key-Value Database

Characteristic	Relational Database	Key-Value Database
Optimal workloads	Ad hoc queries, data warehousing.	Web-scale applications like social networks, gaming, media sharing, and IoT.
Centralization	Typically operated by a central database management system.	No need for central database management system.
Data model	Requires a pre-defined and well-defied data scheme. Furthermore, all of the relationships are defined among tables, columns, indexes, and other database elements.	Does not use a data scheme. The only constraint is a unique key must identify each data item.
Data access	SQL is typically used for storing and retrieving data	Values are directly addressed by their keys
Performance	Performance depends on central database hardware and database administrators optimizing queries, indexes, and table structures	Performance depends on network hardware and latency. More nodes, decentralization, and redundancy improve performance.
Scaling	Relational databases have a hard cap on the number and size of files, imposing upper limits on scalability	No limits on the number of items in a record nor the total size of the record. The capacity of key-value databases scales with the storage capacity of all nodes in the distributed network.

Sources: The Block Research

For modern web applications, the above characteristics confer various benefits:

- **Flexibility** – Unlike relational databases where data structures are predefined as a series of tables containing fields with well-defined data types, key-value databases treat the data as a single collection where every record can contain fields of different types.
- **Performance** - While the primary interface for relational databases is SQL, the structure of key-value databases lends itself to modern object-oriented programming. Values in key-value databases are not represented by placeholders of input parameters as in most relational databases. This feature allows key-value databases to use less memory to store the same database, facilitating performance under intense workloads.
- **Deduplication** – Key-value databases only need to store a given block of data once because each block of data is uniquely identifiable by its hash value. The probability of getting the same hash from different blocks of data is virtually zero (e.g., 2^{-256} for SHA-256).

- **Content immutability** - Writing new data to a key-value database will never change the content associated with a given hash because that hash is unique to that block of data only. Relational databases cannot provide such a degree of content immutability.

- **Efficient caching and preservation** – With key-value databases, it doesn't matter who serves the content or how, so content can be stored and preserved efficiently without dependence on a central database manager.

- **Verifiability** - Any client can verify that a given piece of content is the expected content by recomputing the key for the content returned by a retrieval request.

It's worth noting that AWS offers both traditional relational databases and high-performance key-value databases. That said, the design of key-value databases scales with more node distribution. Therefore, IPFS may gain the upper hand through progressive decentralization and network growth.

Section 5: Incentives & Network Demand

Network demand for decentralized storage networks is directly related with their incentive structures. For example, financially incentivizing making storage available for future deals tends to increase the supply of storage capacity, while financially incentivizing storing data for valid clients tends to increase used storage. The fees for storing and retrieving files are deciding factors when clients choose which cloud to keep their files in.

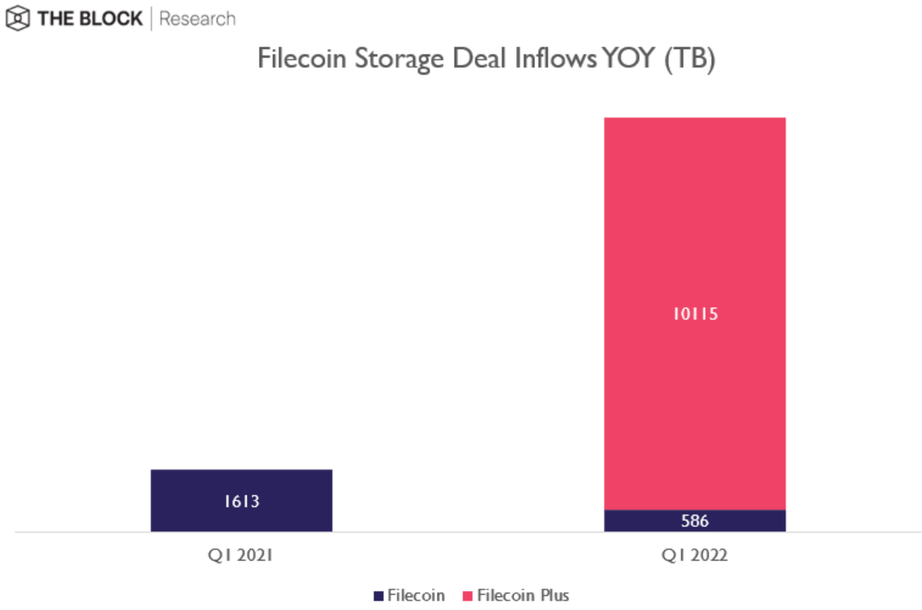
Contract-based decentralized storage networks use similar incentive structures where storage suppliers are paid for storing and retrieving files. There are some slight variations to this general structure, as shown in the table below. The table below also shows how permanent decentralized storage networks like Arweave use a very different incentive structure to support permanent storage.

Incentive Structures for Decentralized Storage Providers

Protocol	Incentive Schemes
Filecoin	Storage entities earn FIL for making storage available and storing client data, retrieval entities earn FIL for retrieving data, and clients pay FIL for storing and retrieving data. Storage contracts specify the storage duration, number of replications, and associated costs.
Filecoin PLUS	One problem with Filecoin's original incentive scheme is that storage providers started abusing it by self-dealing to increase their chances of winning mining rewards. Although this greatly increased Filecoin's commmited storage capacity, quality committed capacity should only include miners matching with real clients. To increase the quality of committed capacity, Filecoin launched the Filecoin Plus program to incentivize storage providers to participate in verified deals with real clients by increasing their block reward share for these deals. Filecoin Plus is a social trust layer built on top of Filecoin to incentivize real storage deals while making self-dealing unnecessary and economically irrational.
sia	Storage entities earn SC for storing data and can set their own fees in SC for retrieving data. Storage contracts specify the storage duration, number of replications, and associated costs. Additionally, storage entities pay a ~3.9% tax on their SC collateral as contribution to the Sia development fund.
STORJ	Storage entities earn STORJ for storing data and retrieving data. Storage contracts specify the storage duration, number of replications, and associated costs. Storage entities are all also paid STORJ for any egress bandwidth involved in network maintenance.
arweave	Arweave doesn't use time-based storage contracts to try to combat malicious actors that withhold, manipulate, or delete client data. Instead, it uses a unique incentive model where clients pay a one-time, up-front fee in AR to store their data "permanently." That is, ~86% of client fees are not immediately distributed to storage providers – instead, they are distributed over time from Arweave's endowment to incentivize persistent storage.

Sources: The Block Research

Incentivizing verified deals over unverified deals a la Filecoin Plus appears to be a useful strategy for growing decentralized storage networks. After the launch of Filecoin Plus, deal inflows flipped from predominantly Filecoin deals to Filecoin Plus deals. Furthermore, the network saw tremendous growth in storage deal inflows, as shown in the charts below.



Source: Filecoin

In Q1 2021, storage deal inflows were ~1,613 terabytes (TB) (~1.6M gigabytes (GB)) on average and virtually entirely driven by Filecoin. By, Q1 2022, that figure rose to ~10,702 TB (~10.7 million GB), or ~564% YOY growth, with nearly 95% driven by Filecoin Plus.

Network Capacity vs. Usage

While the figures above reflect network demand, it's also important to consider network supply – the total storage capacity available across all nodes in the cooperative storage cloud. For any decentralized storage network, network supply must be equal to or greater than network demand for the network to remain viable.

Storage Used vs. Storage Capacity by Protocol

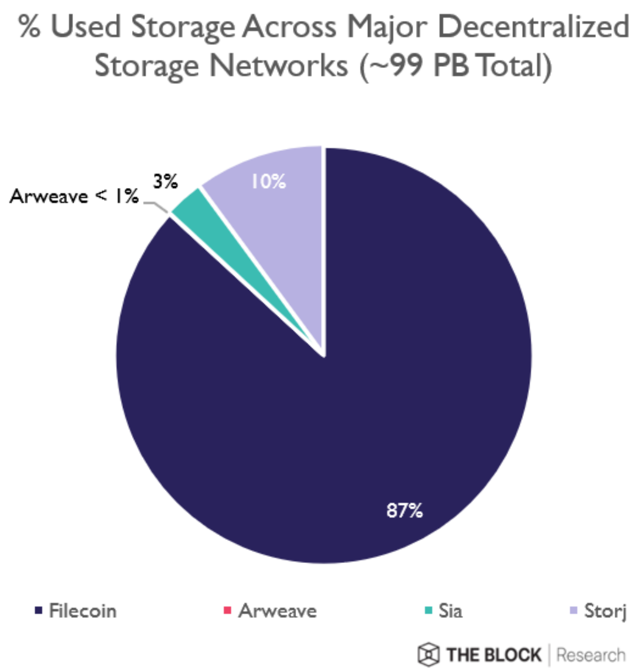
Protocol	Storage Used	Storage Capacity	% Used
Filecoin	~86 PB	~18839 PB	~0.05%
STORJ	~10 PB	~16 PB	~61%
sia	~3 PB	~8 PB	~33%
arweave	~0.06 PB	NA*	NA*

* Arweave's protocol matches supply and demand.
Sources: Filecoin storage used and capacity (2022/04/20); Storj storage used and capacity; Sia storage used and capacity (2022/04/20); Arweave storage used (2022/04/20)

It's clear that Filecoin has amassed the greatest storage capacity by far. At its current size, Filecoin could store about a billion hours of 4K video footage. This massive storage capacity is provided by the cumulative effort of nearly 4,000 active miners comprising its storage cooperative.

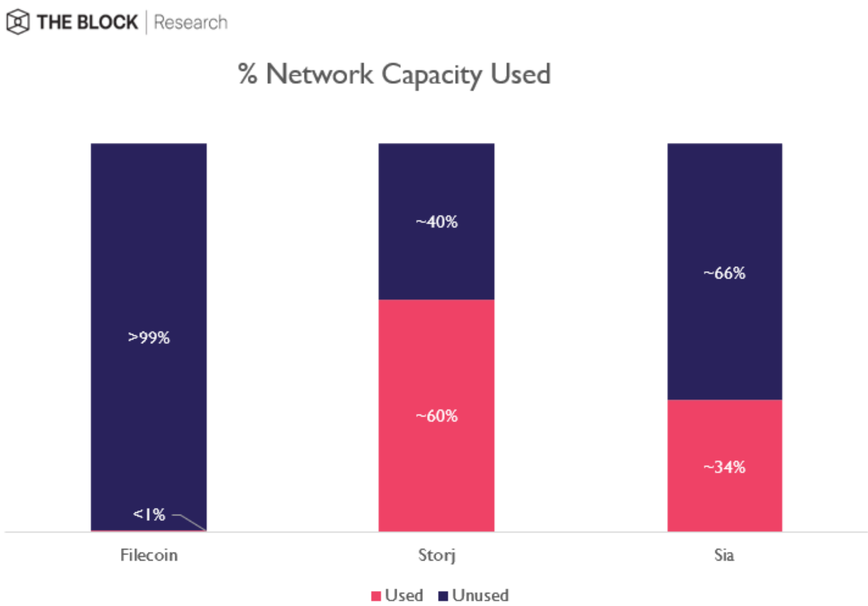
A major driver behind this massive accumulation of storage capacity is Filecoin's incentive system which rewards miners for making storage available to the network in addition to actually storing data for clients. With Filecoin, this incentivized miners to self-deal to prove to Filecoin that they have the capacity for future deals. However, Filecoin Plus tried to deal with this issue by verifying future deals via vetted clients. The intended result is not only more storage quantity but also storage quality.

As for network usage, Filecoin's used storage represented by active storage contracts grew from less than 5 petabytes (PB) (1 PB = 1,000 TB) at the beginning of 2021 to nearly 86 PB as of writing. To put this into perspective, used capacity on competing storage providers Storj was reported as ~10 PB as of March 2022 and Sia reported ~3 PB used storage as of writing. For alternative storage provider Arweave, that figure stood at ~0.06 PB.



Sources: See text, retrieved: 2022/04/20

Now let's visualize the proportion of network capacity used by dividing used storage by network capacity.



Sources: See text, retrieved: 2022/04/20

It's important to take note that Filecoin's used network capacity is relatively less than Storj's and Sia's, but much larger in absolute terms as shown in the figure before. The massive excess supply for future deals makes Filecoin the only protocol so far that is demonstrably prepared to handle web2-scale workloads.

As for competitors Storj and Sia, although their used storage is significantly smaller than Filecoin's used storage, it comprises a much larger percentage of their total network capacity. That said, in all of these cases, there appears to be a reasonable buffer of supply available should network demand suddenly increase.

Section 6: Permanent Storage

While permanent storage is a small subset of the total storage needs of the digital world today, there are not many protocols focusing strictly on permanent storage. Protocols like Filecoin, Storj, and Sia are general purpose solutions that have the capacity for "permanent" storage in the sense of persisting data availability and integrity over long periods of time via external services like pinning and endowment pools.

However, the leading protocol focused exclusively on permanent storage is Arweave. Arweave is combining an incentive-based endowment pool with a blockchain-derived technology for creating a global permanent hard drive.[2] It distinguishes itself from other decentralized storage solutions in several ways, including:

- Storing content directly on a blockchain derivative called the "blockweave." While blockchains are a linear chain of blocks containing transactions, the blockweave is a linear mesh of blocks that connect each block to its previous block as well as a random historical block.

- Securing that content with a novel consensus mechanism called "proof of access" (PoA). PoA is enabled by the blockweave and forces miners to prove that they have access to old data to add new data.

- Using single upfront payment for permanent storage instead of storage contracts for time-locked storage.

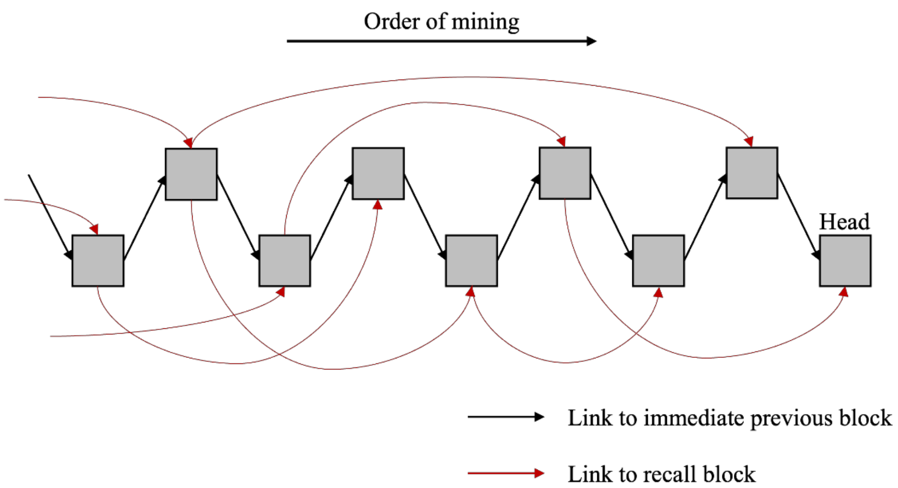
In addition, Arweave has seen the greatest network growth in terms of relative percentages over the past year in both the size of its blockweave and transaction count.

Like the Filecoin blockchain, the blockweave effectively hashes content to make it immutable. Unlike Filecoin, the blockweave stores the hash of the previous block and the content itself. Then, the blockweave is both a storage and incentive layer, unlike Filecoin and other decentralized storage networks that use blockchains only as an incentive layer, keeping data about storage and retrieval deals.

The hypothesis put forth is that the blockweave is better designed for data storage than typical blockchains due to how new blocks are inextricably linked to not only the last block but a random historical block. In combination, all the data required to process new transactions and new blocks is memoized into the state of each block. Therefore, miners do not need to store the entire blockweave. Instead, they are incentivized to store as much of it as possible to maximize their chances of earning mining rewards if they can prove access to the last block and recall block. Effectively, this makes Arweave a state-sharded storage system where no single miner needs to store the entire state, yet the global state is collectively stored by all of the miners.

[2] Though, there are some new protocols like Lighthouse built on IPFS and Filecoin are following the concept of using endowment pools to incentivize permanent storage.

Blockweave Structure

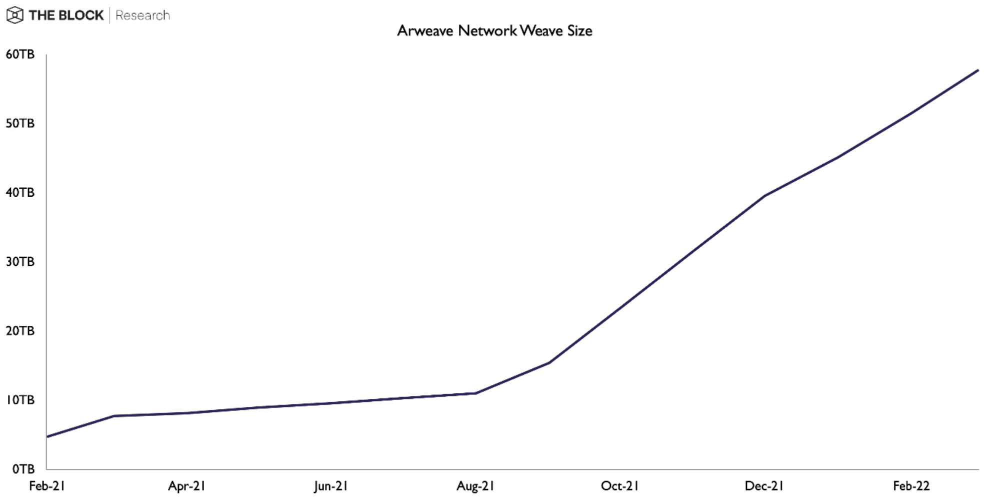


Source: The Block Research

In addition, Arweave uses a unique incentive system where clients don't need to agree on storage contracts with storage providers. Instead, clients send transactions with data and upfront payment directly to the blockweave, and storage providers variably host parts of the blockweave commensurate with their respective storage capacities.

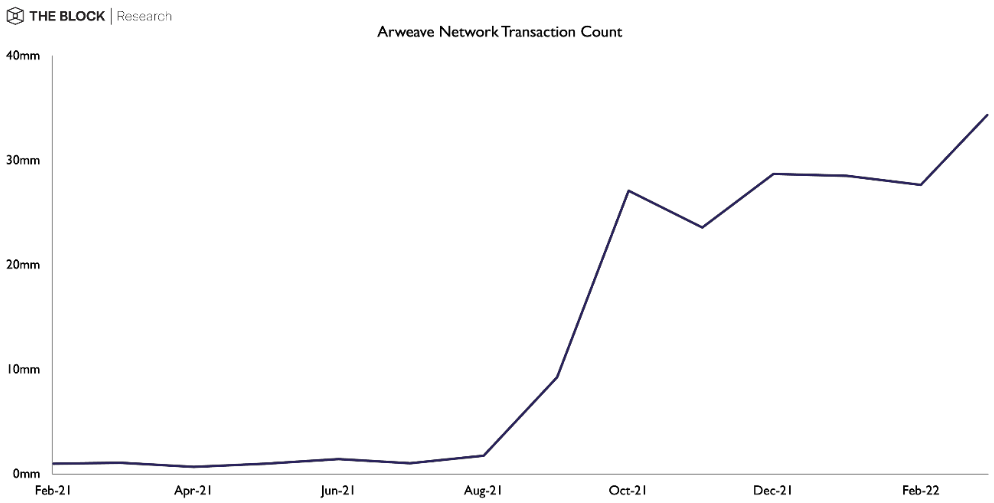
The upfront payment factors in the assumption that advancements in storage technologies will decrease storage costs. Over the past 50 years, hard disk drive (HDD) storage costs have decreased by at about 30.5% per year on average. Taking this and various other variable cost factors associated with cloud storage into account (e.g., electricity, bandwidth, network size, node churn, token price volatility), Arweave offers pricing under the conservative assumption that storage costs will decline by about 0.5% per year. The upfront cost covers the first 200 years of storage at this rate. But, if storage costs decline at a faster rate, the upfront costs can cover many more years.

The technological and economic ingenuity and novelty of Arweave have made it a standout in the decentralized storage market since early 2021. For example, the total size of the blockweave grew from 7.75 TB in March 2021 to 58.55 TB in March 2022, representing a YOY growth of over 650%.



Source: viewblock.io

Total transactions in the Arweave network also accelerated rapidly in August 2021. In July 2021, there were just ~1 million transactions, but by October 2021, that figure rose to ~27 million, and by March 2022, the figure peaked at ~50 million.



Source: viewblock.io

Arweave's ecosystem continues to expand – as of September last year, The Block Research identified 105 projects and companies across 20 different verticals in the Arweave ecosystem. As for Arweave's future, as a new protocol, of course time will tell whether they can uphold their promise for permanent storage.

Arweave may face a major contender for permanent storage in the near future, with Filecoin planning to launch its Filecoin Virtual Machine (FVM) this year. The FVM enable developers to automate permanent storage into Filecoin smart contracts. Both the FVM and Arweave will be able to support data archival and creating a preservation layer for humanity's most important information.

Section 7: Applications of the Decentralized Storage Ecosystem

NFT and Web3 Storage

One of the primary use cases for decentralized storage is storing NFT metadata and hypermedia. NFT metadata refers to a JSON document that includes descriptive information about an NFT, such as its name, what it is about, a link to the associated hypermedia, traits, and so forth. NFT hypermedia refers to the graphics, audio, and video representing digital art, profile pictures, collectibles, music, and so forth that people are paying to own when buying NFTs.

By early 2021, it became apparent that NFT adoption was reaching the mainstream. However, because blockchains are not designed to store large data files like NFT metadata and hypermedia, many such files were not being stored in a decentralized and secure way as people expected. To secure the content and value of NFTs, Filecoin launched NFT.storage in April 2021.

NFT.storage utilizes Filecoin and IPFS to provide a simple solution for decentralized storage for NFT metadata and hypermedia. It aims to provide secure, immutable, and perpetually redundant storage through IPFS servers currently hosted by Filecoin. Because IPFS is a standard used by many different storage services, it's easy to store data uploaded to NFT.storage on any IPFS-compatible storage services, including remote and local pinning, as well as other decentralized storage networks like Arweave, Storj, and Sia.

Since its launch, NFT.Storage already secures nearly 65 million NFT uploads totaling nearly 170 TB of data. Its user base extends from individual artists to emerging and large NFT marketplaces, including OpenSea, Holaplex, Magic Eden, MakersPlace, Metaplex, and more.

Similarly, Web3.Storage is another hosted storage service launched by Filecoin soon after NFT.Storage for storing any kind of data for web3. According to its site, Web3.Storage has already stored data for over 2,500 web3 projects, comprising over 15 million data objects stored redundantly across over 60 Filecoin storage providers.

The long-term goal of Web3.Storage is to become the de facto storage layer of web3. To that end, Filecoin is building its partnerships with major protocols in the web3 space, including Polygon and Solana. In August last year, Polygon announced its native bridge to Filecoin, envisioning an interoperable environment where Filecoin brings greater functionality to Polygon applications that require decentralized and verifiable data storage services. Filecoin has also seen a lot of traction in the Solana ecosystem lately, with various Solana NFT marketplaces including Metaplex, Magic Eden, and Holaplex offering native support.

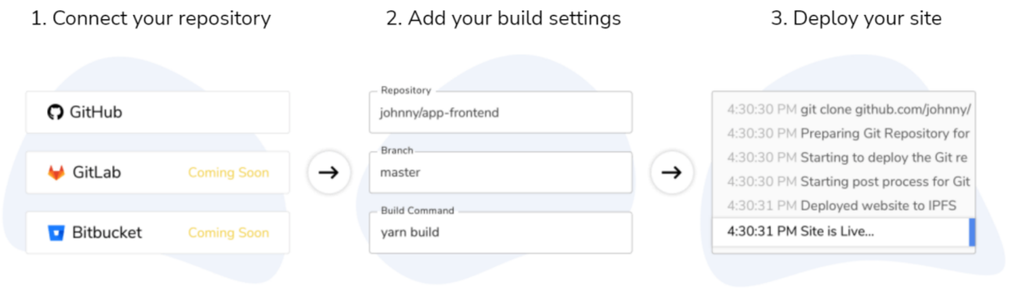
NFT.Storage aims to leverage IPFS and Filecoin to create a public commons for NFT data – making storage and access free for this shared data asset. When Filecoin launches smart contracts later this year, the plan is to form a DAO to bring together a community to manage on-chain endowment to perpetually fund this public good, with revenue generated by on-chain activities like loans and DeFi. However, with Web3.Storage decentralizing, clients can expect to pay for storage in the future – either directly through fees or indirectly through sharing staking rewards.

Dev Tooling

Another area rapidly developing in the decentralized storage ecosystem is developer tooling and the communities forming around them. From integrative development environments to web hosting to various services for facilitating file and code storage and usage, it is becoming clear that developing on decentralized storage networks is rapidly streamlining:

- **Development Environments** – Textile builds a wide range of dev tools for Filecoin including Deal Auctions for streamlining the storage-deal-making process; Storage Bridges for connecting Filecoin to other Layer 1 and Layer 2 blockchains through permissionless bridges; Buckets for allowing developers to use IPFS like traditional cloud storage via dynamic folders stored locally and pinned on the IPFS network; and Hub and Powergate for facilitating the onboarding process for new IPFS developers. Textile's new and popular product Tableland can be used for relating and extending assets stored on Filecoin via familiar mutable relational databases (vs. key-value databases natively used by Filecoin) – allowing developers to "build web3 with SQL."

- **Website Hosting** – Fleek is a service that lets people host their websites on IPFS without needing to install anything on their computer or deal with the command line. It solves a problem of content addressing which is that changing even a single character to any file will create an entirely new hash for that file. Fleek offers a simple workflow where once the dev pushes their changes to Git, Fleek will build, pin, and update the site for them automatically.

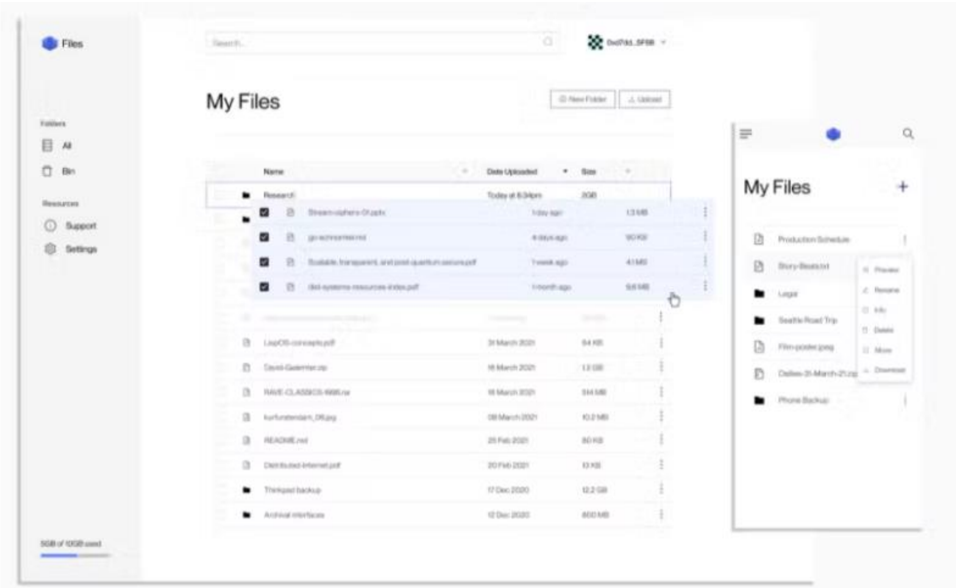


Source: Fleek

- **Automated Storage** – Estuary is a simple IPFS node that automates storage activities on Filecoin via a minimal Filecoin client library. It also repairs lost data and guarantees data replication by replicating the data six times. The accessibility of its interface has made it a popular way to store data on Filecoin. As of writing, users of the Estuary node have pinned over 16M files and directories (175+ TBs) to IPFS and secured nearly 100,000 storage deals.

- **Code Storage and Preservation** – Truffle is a development environment, testing framework, and asset pipeline built for any EVM-compatible blockchain that offers native support for preserving files and directories to IPFS. On Arweave, Gitopia is a decentralized code collaboration platform with built-in functionality to store and preserve versioned code to the blockweave.

- **File Storage, Viewing, and Sharing** – ChainSafe Files, built on ChainSafe Storage, is a Dropbox-like solution for Filecoin and IPFS. The frontend enables users to easily store, view, and share files on IPFS, following intuitive design principles of successful traditional cloud storage like Filecoin and Google Drive. ChainSafe Storage makes it easy for developers to connect to, pin, and built on IPFS and Filecoin.



Source: ChainSafe Files

Emerging Use Cases

In theory, decentralized storage should support any use case that centralized storage can, while offering some unique advantages such as immutability, permanent storage, and transactional tracking that could make it a more attractive option depending on the nature of the app's storage functions. Looking at the emerging decentralized storage ecosystems, the possible applications of decentralized and persistent storage are numerous and diverse, including:

- **Metaverse/Gaming** – Decentralized storage can be useful for user monetization in metaverse games. For example, in Mona, gamers display their digital artwork in virtual art spaces and sell them to collectors. Any transaction transferring data from one player to another can be tracked on the Filecoin blockchain. In addition, Blast and Gala Games use Filecoin for backup data supporting gamers' payments and revenue generation.

- **Social Networks/Communication –** Matrix is a communications protocol built to support chat, VoIP, IoT, VR/AR, social, and more. Originally a server-oriented network, Matrix is evolving into a hybrid P2P network with IPFS supporting the P2P Matrix. The vision is to empower users to have more autonomy and privacy over their data (e.g., by storing the data in IPFS by embedding their own servers into their Matrix client). Matrix powers Element, a Matrix-based messaging app. RSS3 is an open protocol designed for content and social networks in web3, with integration with Arweave.

- **Content Publishing –** Interactions with decentralized storage networks that leave a permanent trail of blockchain (or blockweave) transactions provide information that can be used for new content monetization strategies. For example, the Koii protocol enables people to monetize attention on any content (e.g., NFTs) stored on Arweave by tracking interactions with that content.

- **Content Delivery Networks** – Myel is building a community-powered decentralized content delivery network (dCDN) powered by Filecoin and IPFS. In the dCDN model, clients are servers and servers are clients, meaning that as more users use an app, its infrastructure scales. Users of an app are simultaneously the backend servers, points of presence, and customers for the app. Thus, apps can scale without relying on expensive server-side infrastructure. Meson Network is an example of a decentralized CDN powered by Arweave.

- **Permanent Storage/Web2 Datasets** – Similar to Arweave, Lighthouse.Storage is a new service that allows Filecoin users to pay an upfront fixed fee for permanent data storage via a storage endowment pool. Such services can be useful for securing the content and value of NFTs or for data archival. Furthermore, large public datasets like the Internet Archive use Filecoin to archive websites and documents for public access. On Arweave, projects like Akord and ArDrive are providing similar services.

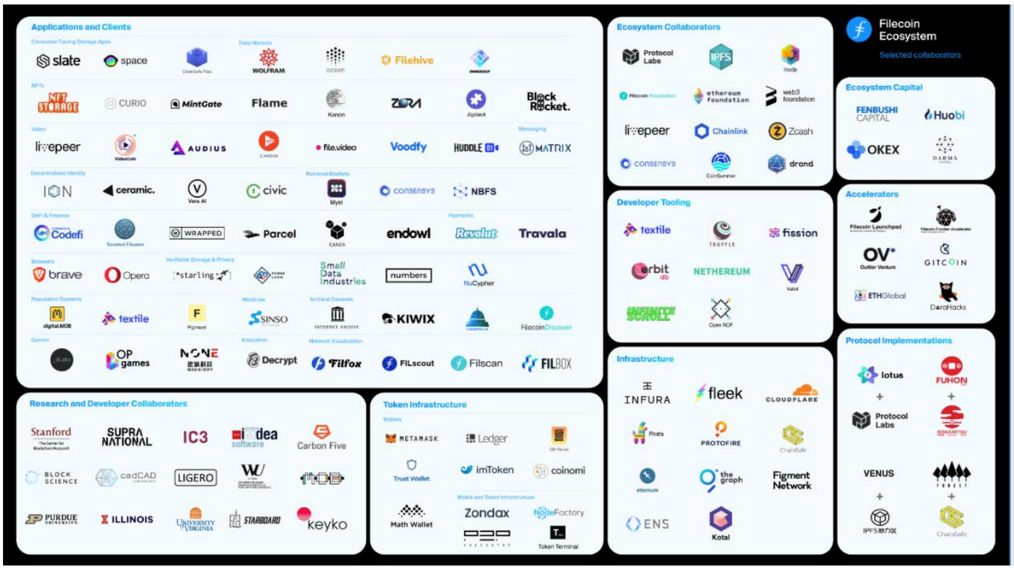
- **Audio/Video** – Audius is a prominent music streaming service like Spotify that uses IPFS for storing music files. Huddle01 is a video conferencing platform like Zoom that uses IPFS for storing recorded meetings. Its current focus is remote meetings for NFT communities. On Arweave, Bandplay is building a music marketplace where users can transact music using Arweave's native token.

- **Decentralized Identity** – ION is one decentralized identity solutions using IPFS for storing identity files and proofs. By doing this, ION can offload a significant amount of its storage requirements as the ION network only needs content identifiers from IPFS to gather identification data about any entity within its network.

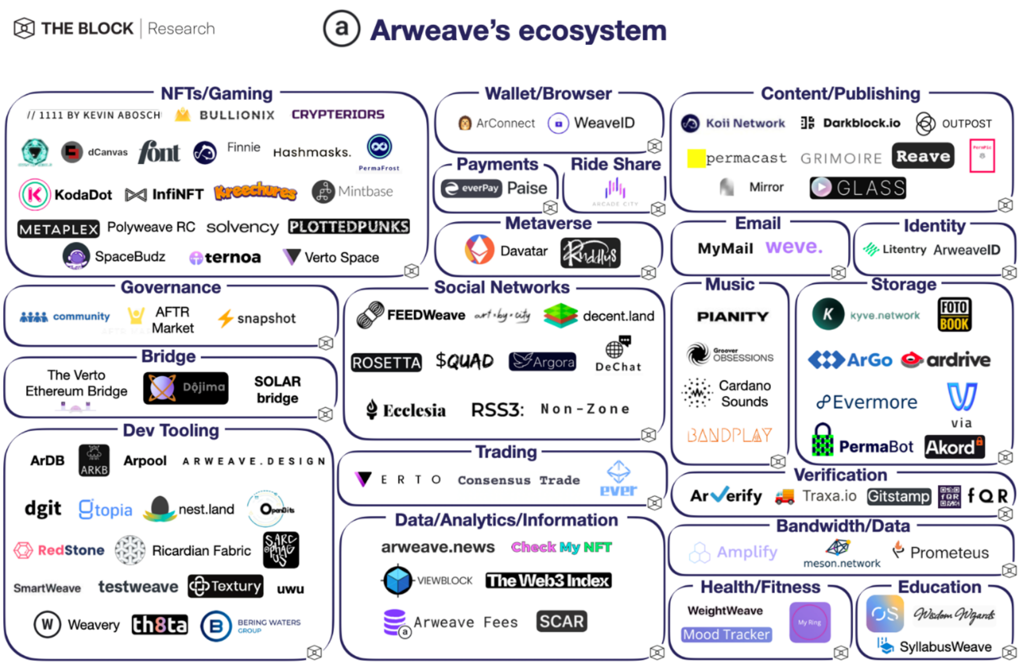
- **Computation –** Ceramic is a decentralized content computation network. It is built upon IPFS and features a permissionless design that allows anyone to openly create, discover, query, and build upon existing data without needing centralized servers, one-off APIs, or worrying about data integrity owing to IPFS data integrity controls. It is compatible with persistence networks, including not only Filecoin but also Arweave and Sia.

- **Profit Sharing Tokens –** Profit sharing tokens (PSTs) and their communities allow developers to earn a stream of micro-dividends while users interact with apps they built on Arweave's permaweb. Every interaction that produces an Arweave transaction will generate a small profit for PST holders. Verto is one decentralized exchange where people can trade PSTs to exchange their ownership of a share of the profits generated by apps valuable enough for users to pay for interacting with them. PSTs are a unique feature available to apps stored on Arweave, as they are stored on the blockweave itself instead of on node servers.

By looking at the ecosystem maps for Filecoin and Arweave below, we can start to see that use cases and projects associated with decentralized storage are variegated and burgeoning. Both ecosystems rapidly expanded from a few dozen or fewer projects at the start of 2021 to over [330]() projects on Filecoin and over [100]() on Arweave today.



Source: Filecoin



Source: [The Block Research]()

Section 8: Conclusion and Outlook

Blockchain- and crypto-enabled decentralized storage is off to a promising start, with the rapid adoption in 2021 showing promise for the tech and incentive systems powering decentralized storage networks. The leading contract-based decentralized storage protocol Filecoin and the leading permanent decentralized storage protocol Arweave have rapidly developed lively ecosystems that point to the possibility of a "sub-web" running on P2P servers instead of centralized databases.

The lively communities supporting these ecosystems are likely to keep developing new features for these decentralized storage networks. On the technological front, there are many opportunities to bolster these systems through cutting-edge encryption, redundancy, and communications technologies.

We must wait for greater adoption to see how sustainable, scalable, and reliable such systems are and whether they can power the memory-intensive web apps of the future. The technologies powering decentralized storage and their advantages suggest they may achieve this feat – for example, low-latency content addressing, object-oriented key-value databases, and zero-knowledge proofs and economic games for providing trustless service guarantees.

Although blockchains are generally not made for storing a lot of data, Arweave's blockweave is showing that it may be possible to derive blockchain-like databases by modifying path designs and consensus mechanisms.

The success of these protocols depends on the applications built upon them. So far, NFT and web3 storage are the prominent use cases, but emerging new applications in areas ranging from audio/video to gaming to computation and more are making the case for global decentralized storage networks accumulating more of the cloud storage market.

Section 9: Disclosures

This report is commissioned by W3BCLOUD™. The content of this report contains views and opinions expressed by The Block's analysts which are solely their own opinions, and do not necessarily reflect the opinions of The Block or the organization that commissioned the report.

The Block's analysts may have taken positions in the assets discussed in this report and this statement is to disclose any perceived conflict of interest. Please refer to The Block's Financial Disclosures *page for author holdings. This report is for informational purposes only and should not be relied upon as a basis for investment decisions, nor is it offered or intended to be used as legal, tax, investment, financial or other advice. You should conduct your own research and consult independent counsel on the matters discussed within this report. Past performance of any asset is not indicative of future results.*