Review of Data Storage Techniques

Sejal Jain
Computer Department
Dr D.Y Patil Institute of Technology Pune, India
sejalj519@gmail.com

Anil Kumar Gupta
Senior Member IEEE, CDAC
anilkgupta@ieee.org

Hardeep Kaur Ruprai
Computer Department
Dr D.Y Patil Institute of Technology Pune, India
hardeepk.1074@gmail.com

Hardeep Sinha
Computer Department
Dr D.Y Patil Institute of Technology Pune, India
hardeepsinha21@gmail.com

Amarjeet Sharma
CDAC
amarjeets@cdac.in

Priyanka Ghosh
Computer Department
Dr D.Y Patil Institute of Technology Pune, India
priyankaghosh0506@gmail.com

***Abstract* - With the massive increase in data, there is need for data storage systems which are maintainable, accessible, secured and affordable. Large scale storage systems often give importance to features like reliability and availability. Such features have become desired and essential for any storage system today. Storage methods can be Centralised or Decentralised, Cloud-based or Non-Cloud based depending on requirements of storage. Introduction of Blockchain is enhancing the functionalities of the storage methods. Blockchain primarily targets the security and credibility of data, which are most important in today's times. Blockchain gives an added advantage of security to the data which majority of the systems fail to deliver. The recent projects Sia, Swarm, and Storj, have decentralised distributed storage formed on Blockchain and can revolutionize the storage system's view. This paper evaluates the various types of storage systems; compares them based on their centralised, decentralised and distributed nature and sheds light on blockchain and how its utilisation can be viewed with varying storage methods.**

Keywords— Blockchain, Cloud-Based Storage, Centralized Storage, Decentralized Storage, Non-Cloud Based Storage, Sia, Swarm, Storj, Rootstock

I. INTRODUCTION

Nowadays, data is more valuable than money; it is essential to treasure data because it contains sensitive information, transaction records, financial records etc. Storage systems that can store the data such that it remains safe and can be accessed quickly and feasibly are required. The evolution in internet and technological fields has brought changes in the data storage functions and gave a different view on the data storage systems that explore centralised storage systems' fallibility in contrast with decentralised storage. Many emerging storage systems now use blockchain to make it more secure, effective and the one that can give easy access to the storage system. [1]

Earlier file sharing and storing data was basic and rudimentary. But the evolution of technology has made it possible and easy for users to share files and access data from anything and anywhere. Now there is an increased usage of cloud services like S3 that provide scalable hosting. In such systems, users hand over their data to the third party and its security and privacy becomes beyond the user's control. Centralised storage can be expensive. In decentralised storage, the user's files are split across multiple nodes throughout the network. Hence, it ensures that retrieving the data won't be so challenging if it encounters any failure. Nowadays, blockchain has become the hottest use case of decentralised storage. Sia, Swarm, Storj are some examples of decentralized storage projects. [1]

Blockchain Storage: It is a network of cryptographic blocks linked by hash, and all transactions are monitored over the blocks sequentially. It uses a simultaneous algorithm that ensures data consistency through the nodes.[2] The information stored in the blockchain becomes permanent and secure when it becomes the time-stamped encrypted data log network. It also guarantees accountability and needs zero operating procurement of all infrastructure.[3] The blockchain can be either centralized or decentralized; it relies on the ledger's participants' interests.[4] In comparison, the paper is typically broken into centralized, decentralized and distributed database framework structures. Besides, the subsection is broken down into cloud-based and non-cloud-based storage; Section III is about decentralized storage and cloud-based Storage over the blockchain. Section IV gives detailed information about distributed storage.

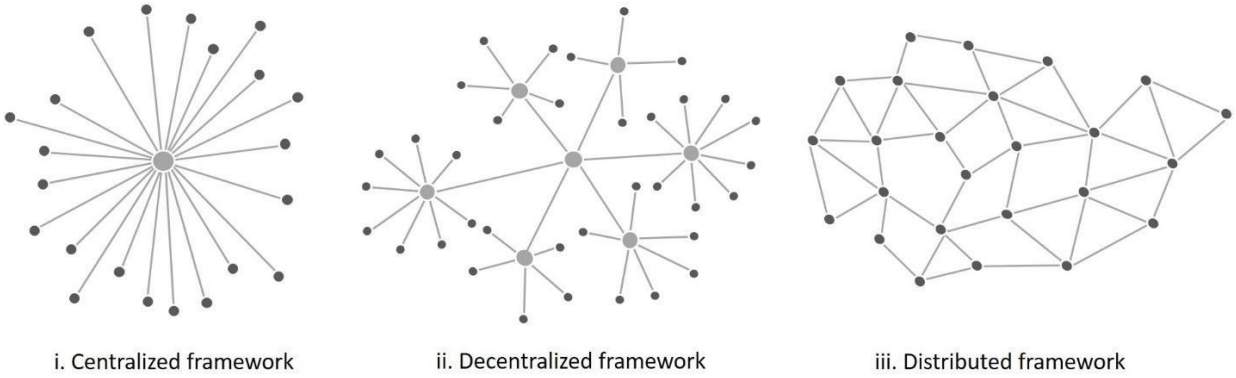

Fig. 1 Storage systems can be divided into centralised, decentralised and distributed storage

II. Centralized Storage

When information is stored, maintained and located at a single location, the storage is known as Centralized Storage. There is enormous potential for centralised blockchains. Though Blockchain is distributed, it is possible to have centralised blockchains as well. But the participants can post to the ledger only if it is a centralized blockchain.[4] Proof of Work is an algorithm used to confirm transactions, and it generates new blocks over the chain. In the Bitcoin crypto network, miners complete the proof of work and regulate the block generation rate. This paper compares the centralised cloud-based storage systems with non-cloud storage in variation with blockchain.

A. Cloud-Based Storage

Cloud services allow storing tremendous data over the cloud. Millions of devices sync and duplicate the subset of their local storage in a central server cluster, and millions of users connect to that server and access their files. In cloud storage, users can have their accounts to the centralised server, and they can associate the files with that account. But such centralised cloud systems have a lot of inherent issues. A significant problem of version control and duplication can arise when users duplicate their files to access them later on the internet. This eventually makes cloud storage more complex. Another issue can be the facilities of managed and secured data [2]. There are not so many techniques to protect data stored over the cloud servers with comfortable management. Some propose to encrypt the files before storing it. BigchainDB is another database used for cloud storage which ensures integrity and deduplication [5]. The blockchain provides the integrity and non-reduplication of information.

B. Non-Cloud Based Storage

Previously non-clod storage systems like hard drives, floppy disks, CDs, etc. were used for storage purposes. These devices had a large space for storing data. They were rudimentary, and the users had to own and maintain their servers [1]. Conventionally there was a need for secondary backup devices to store the data if the storage space was exceeded. This concept is prevalent even today for storage techniques. But connecting different devices and servers over different locations in the world is unachievable for non-cloud storage. In such systems, data cannot be retrieved If lost from the central device. As against the blockchain, encrypted data can be stored in non-cloud storage.

III. Decentralised Storage

Data can be stored via various servers and computers in Decentralized Storage, where files are encrypted with cryptography or blockchain.

Decentralized storage has the capacity to transform storage networks. Data from anywhere and everywhere can be accessed in such systems. The new approach that blockchain firms are introducing is decentralized storage. The idea is to retain users with sophisticated cryptography and encryption versions of files with network members. Decentralized storage has become a swift solution to all the challenges of centralized networks. Sia, Swarm, RootStock, Storj [1] are the most common and evolving decentralized storage projects. Decentralized cloud-based storage can overcome all the issues faced by a centralized cloud. It offers online connectivity and data sharing to all sites, whether in the cloud-hosted network or on-site, instead of transferring and storing data in a single data centre. It is situated a mile away from all connected devices. The Hybrid Architecture FileFlex[8] is decentralized. Following are some decentralised cloud storage systems that are based on blockchain

A. Sia

Sia is a decentralized, cloud-focused storage network. It is a stable peer-to-peer network through which users can communicate. It aims at creating data storage which is safe and cheaper than current solutions[6]. Sia has two parties, mainly the storage provider(host) and client. The clients store data on contractual basis. It leases storage and blockchain is used for that purpose. For a given time, there are arrangements between storage vendors and consumers. Sia is a peer mechanism in which the host can be held liable for proof of work; the host can even be penalized for lack of proof. There is proof that the file contract must be satisfied. Sia uses a multi-signature M-of-N mechanism for all transfers as opposed to Bitcoin. By using contracts, proof, and contract changes, it allows storage contracts. Contracts provide the concept of file storage size and hash for the host, and the host shows credence of the regularity of the storage proofs. The contract changes shall be used to amend the contracts entered into. Per entry is authenticated in a transaction with a cryptographic signature. Each signature is combined with a particular input identification. The clients and hosts in sia are sometimes vulnerable to different attacks like the Closed window attack, Sybil attack, Block Withholding attacks. Various precautions are also taken to reduce such attacks, but that does not guarantee an entirely attack free system. The clients and hosts are provided with multiple protections to minimize the possibilities of such attacks.

It is possible to send Siafunds like Siacoins to a different address by imposing a fee on such contracts. 3.9% of the contract fund is removed and combined with the Siafunds to generate revenues for Nebulous Inc. [7]

B. Storj:

It is open-source, decentralized cloud storage built over ethereum. Storj preserves the data via a distributed hash table, sharded and encrypted in a secure area. It has many advantages such as reduced costs, secure and private storage, and increased scaling. It is S3-friendly.

The storj is similar to the technology from Torrent, which was popular at the beginning of the 2000s. Data are first encrypted and then exchanged in Storj (i.e. information is split into smaller pieces). These encrypted shards are then saved over the storj network redundantly. When the user accesses a file, he discovers all files and works them with hash tables. Kademila, a distributed hash table which is mostly being used for this reason. Both files can only be encrypted and read by their users until they are uploaded. If one of the nodes collapses, files can easily be accessed. The storj conducts audits to ensure files are available.[10]

In the Storj, there is also a Storj Lab, a body which joins the business with decentralized cloud storage. The rental is paid, and the network is rented to customers. Storj supports Daas (Data-as-a-service)

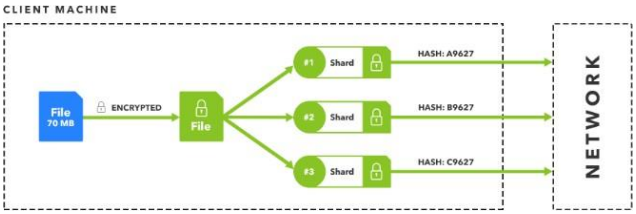

Fig .2 Sharding Process.
Source: Adapted from[11]

C. Swarm:

For decentralized storage, Swarm is often used. A variant of ethereum for the decentralized Network is the Swarm Protocol. It operates along with whisper[6] The programs which are offered by supplying each other with tools. It is focused on smart contracts for financial urgency. It is a peer to peer network, and both will connect to the network through the swarm client. It operates on all platforms. It is written in golang and needs to run geth (go ethereum client). The Swarm Network with its Network ID is established. There is a chunk, guide and manifest the three pieces of the architecture. The key storage portion in the swarm is known as a chunk. The client file has a special reference identity that allows data recovery and access. This is a hashed cryptography of the data which acts as the address of information. The file set explains the Manifest. Manifest can also be used by mapping files into the file system file tree to upload and import. Manifest can also be indexed for fundamental hosting pairs and includes a virtual hosting service. The section of the reference addresses that should be collision-free, deterministic (if the information stays the same, then the identifier would still be the same), and uniformly distributed, should be one identity block.

ENS (Ehereum Name Service) is an intelligent contract that permits domain owners to register a link to their domain. ENS offers a mutable resource facility. Each node has its own base address from the Account Public Key (Keccak 256 SHA3 bit) hash. Swarm includes an HTTP proxy-based API for communicating with DApps. The command is issued in order to store a node file, and if the file is uploaded, a hex string is generated as output. This hash is appropriate to download the file called swarm hash[9]. The submitted content is broken into chunks, and each chunk is accessible with a hash across each chunk. The referrals to the chunk are then bundled into a chunk with its own hash. It permits chunks to be seen by the Merkle. When published to the Swarm, the data cannot be revoked[9].

D. RootStock

RootStock(RSK) is a platform associated with bitcoins' blockchain and is a smart contract-based. It has a technology stack RIFOS Infrastructure Framework Open Standard. RootStock is constructed on top of a side chain(RSK) that resides on top of Bitcoin. RIF storage is a protocol bringing solutions for decentralised storage solutions. Without the need

for a server, it can be used for DApp. Access to several storage systems Sia, IPFS [6] can be done by RIF storage.

In the following table, the pros and cons are discussed.

Though Systems like Sia, Swarm and Storj are decentralised, they have their advantages and disadvantages.

TABLE 1
PROS AND CONS

Technique	Pros	Cons
Sia[7] (cloud-based)	1. Peer to peer system 2. Secured 3. It does not have a single point of failure. 4.Affordable as compared to other storage options 5. There is automatic verification of storage proofs; therefore, users need not verify the storage proofs personally. 6. It is a storage network based on contract - this helps to establish confidence for the customer that the data is stored	1. The host could be fined for missing proof 2. There is no practical method for customers to choose quality hosts, hence the clients have to be very careful before choosing a dedicated host. 3. There can be chances of Sybil attacks 4. Hosts are prone to closed window attacks
Swarm[9] (cloud-based)	1. Peer to peer system 2. Allows functionality of virtual hosting 3. Swarm runs on all major platforms. 4. Supports encryption 5. Files uploaded on Swarm are first broken into chunks, encrypted, then hashed and stored.	1. Data cannot be revoked once uploaded(unable to delete or remove). Hence users should avoid uploading unethical, illegal or controversial data. 2. There are risks of important content being purged: if the nodes' storage capacity is reached, The garbage collection system eliminates the least accessed chunks. 3. Swarm cannot be regarded as a completely safe storage platform
Storj[10] (cloud-based)	1.Peer to peer system 2. Open Source 3. Secure 5. Makes cloud file storage faster 6. It provides high user privacy as it is end-to-end encrypted 7. It has lower costs (affordable to use)	1. Prone to attacks like Sybil, Spartacus, Eclipse, Hostage bytes.

IV. Distributed Storage

The term ‘distributed ‘ means the difference of locations and distributed storage is the type where data is stored at distinct locations. Blockchain is the application of distributed storage. Unlike decentralized storage, all the system parts are not in one physical location. The system's components are placed at distinct locations. All nodes are connected to, and one node's failure in a distributed system will never fail the entire storage system. Data can be easily recovered from another location. Here are some of the distributed storage projects:

A. DAOS

DAOS, an open-source object storage designed for non-volatile memory that is massively distributed. For the optimization part, transactional I/O(non-block), automatic data protection, end-to-end data confidentiality, data management and elastic-storage, cost and system performance are some features which are optimized and help it to stand out.[12]

B. MooseFS

MooseFS is a distributed framework with high resistance to faults. It is a POSIX file system. Because the filesystem is distributed, the stored user data is scattered all around diverse locations globally and are stored over servers. It has four components: application administration, data server, metadata backup server and client computers. Managing servers keep the 2track of the metadata stored over the server. 1The data stored is then broken down into chunks and stored onto chunk servers. It has an MFS installation phase and then interacts through the network. For clients, the operation of the device is transparent. It can be run on all operating systems. The machine has unique needs 2such as a TCP/IP network. Chunk

servers, master servers and CGI servers also have different requirements. [13]

Vasto stores on the system. All the clients are manoeuvred by the master(the brain of the system). To link to the correct store across each client, it relies on the master. Vasto stores are nothing but a simple wrapper of RocksDB. There are several gateways in Vasto which supports numerous APIs using client libraries. Gateways to Vasto are limitless. For allocating data, the device uses a Jumping consistent hash. One Vasto cluster has one master, and when the store enters the cluster, most stores are vacant. Masters are asked to build and resize

C. *Vasto*

Vasto can be described as a key-value store that is distributed. There is one Vasto master, and others serve as keyspaces. Vasto is an in-house cloud that as a service offers distributed key-value stores, minus the need to balance the expense of performance and cloud service. It provides stable and low latency.[14]

The following table compares the advantages and disadvantages of the distributed storage systems mentioned above.

TABLE 2
PROS AND CONS

Technique	Pros	Cons
Daos[12]	1. Open Source 2. It has high bandwidth and low latency 3. Focuses on providing high IOPS (Input/Output Operations Per Second) 4. Provides end-to-end data integrity.	1. It does not support Disk-based storage as it is designed particularly for SCM and NVMe.
MooseFS[13]	1. MooseFS is open source and is best for applications that require high performance. 2. It is a Fault-tolerant file system. 3.It offers high availability. 4. It supports I/O operations that require high performance. 5. It has a highly scalable storage capacity and can store more than 2 billion files. 6. Deployment and maintenance is easy. 7. It provides many features like recycle bin function similar to the one in Windows, Garbage collector similar to the one in Java. 8. It provides support for Big Data applications. 9. It has no single point of failure	1. The Master Server faces numerous issues due to bottleneck, which affects the overall performance of the system. 2. The Master Server requires more memory as the number of files increases. 3. The Metalogger server takes longer time to copy the metadata. 4. The Master server can face issues due to single point of failure
Vasto[14]	1. It is a key-value store. 2. This distributed system offers eventual consistency. 3. Vasto uses an active-active Replication system to perform replications in just milliseconds. 4. To perform fast failure detection, fast topology changes, and for error-free co-ordinations, Vasto requires only a single master. 5. It is very easy to recover the Vasto master from crashes if any. 6. There can be an unlimited number of Vasto gateways/proxies. 7. It has High Availability.	1. The HDD storage capacity is very large but it has limited IOPS. 2. The SDD/Flash has a good IOPS but has limited space. 3. It is difficult to reduce the write IOPS.

V. CONCLUSION

This paper studies different centralised storage, decentralised storage and distributed framework solutions. Centralised storage has several drawbacks. The charge for storing data in a centralised system is very high, and it seems to be only increasing. The centralised storage offers low security, and there can be chances of data leaks. The speed of data

transmission is relatively slow because several centralised servers are placed in remote areas. The reasons mentioned above make centralised storage a less preferable choice for data storage.

In contrast, decentralised storage overcomes a few of the disadvantages faced by centralised storage. There is no single point of failure in decentralized systems, which makes them more reliable. By eliminating the intermediate server and storing multiple copies on different nodes the provide advantages like low storage costs, high privacy, faster speed and greater security. Distributed systems also offer certain benefits like high fault tolerance, reliability, high performance, etc.

VI. FUTURE WORK

The systems like Swarm, Sia, and Storj can be made even more reliable by reducing the attacks. As there are advancements being made in the Blockchain technology and with the ever-increasing demand for storage, the storage systems in the near future will largely be relying on decentralised and distributed solutions. With the help of Blockchain, decentralised applications that are transparent and resilient can be created. Hereafter the demand and usage of decentralised and distributed applications for storage will grow exponentially.

REFERENCES

[1].https://www.forex.academy/centralised-vs-decentralised-storage-how-blockchain-is-redefining-data-storage/

[2] : Yugala based Encrypted cloud storage for IOsT data 2019 IEEE International Conference on Blockchain.

[3]https://www.techfunnel.com/information-technology/blockchain-storage/

[4] https://www.finra.org/sites/default/files/2017_BC_Byte.pdf

[5] I. Sukhodolskiy and S. Zapechnikov, "A blockchain-based access control system for cloud storage," *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Moscow, 2018, pp. 1575-1578, doi: 10.1109/EIConRus.2018.8317400.

[6]medium.com//what-is-decentralized-storage-ipfs-filecoin-sia-storj-swarm-5509e476995f

[7] https://sia.tech/sia.pdf

[8] https://www.thechannelco.com/sites/thechannelco/files/1045_qnext_-

[9] https://swarm-guide.readthedocs.io/en/latest

[10] https://coincentral.com/storj-beginners-guide/

[11] https://cryptonews.com/coins/storj/

[12] https://daos-stack.github.io

[13] https://moosefs.com/Content/Downloads/moosefs-3-0-users-manual.pdf

[14]https://awesomeopensource.com/project/chrislusf/vasto