Decentralized Storage - Literature Review Jonathan Healy

Decentralized storage is a concept that has roots in peer to peer file sharing technologies of the past. It has emerged as a response to modern cloud computing and naturally shares the ideas and philosophies that helped drive the early development of Bitcoin. The five decentralized storage solutions covered in this review share many things in common. The first and maybe most obvious is that each has adopted the idea of using a blockchain to secure a digital currency in order in incentivize their ecosystem.

The core logic behind decentralized storage is relatively simple. Users contribute hard drive space to store other Users files and are rewarded for doing so. The challenges involved with implementing this model however, are not so straightforward. One common approach is to implement a solution, like the one presented in the paper, “Sia: Simple Decentralized Storage” (1), where a blockchain is utilized to not just store payment history but also to track and activate smart contracts between data providers and subscribers. Data itself is not stored on a blockchain as doing so would be prohibitively expensive and the system would suffer from serious latency issues.

Another issue that is of concern in decentralized storage models is in auditing the network. If nodes are being rewarded for storing data or are under contract to do so, how can the system ensure that they have been? David Vorick and Luke Champine in their paper, “Sia: Simple Decentralized Storage” (1) take the position that contracts themselves should, “require the storage provider to prove, at regular intervals, that they are still storing their client’s data.” In the paper, “Storj: A Decentralized Cloud Storage Framework” (2) a similar idea is used based on probabilistic per-file audits called ‘proofs of retrievability’. Storj also introduces the idea of using a reputation system to rate storage providers.

Filecoin, presented by Protocol Labs in the paper, “Filecoin: A Decentralized Storage Network” (3), claims to introduce a novel proof-of-storage scheme called ‘proof-of-replication’. They also have developed a consensus model based on sequential ‘proofs-of-replication’, using storage as a measure of power called, ‘proof-of-spacetime’. The team behind Filecoin previously developed IPFS, the Interplanetary Filesystem, which is in use today, “serving billions of files across a global peer-to peer network.” Filecoin essentially acts as an incentive layer on top of IPFS. They present 2 verifiable markets, a storage market and a retrieval market. These markets run on top of a blockchain which is similar to Bitcoin except that storage providers will essentially be minting new tokens by providing storage resources. The chances that any one entity has to hash and create a new block (and receive the reward) will be proportional to the amount of storage they have being used in the network.

The need for auditing data on a network is important for many reasons. Nodes storing data may be unreliable and unfit to be participants in the network or they may be looking to find ways to game the system and receive profit for data that they are not sufficiently making available. If, after being audited, a node is deemed as being unreliable or incapable of hosting data, the network must redistribute the data it is holding to new storage nodes. This introduces a large amount of complexity. Contracts need to be revoked, other contracts need to be created, and pieces of data must find a new home and be accounted for.

Node churn, where nodes join and leave the network, is a related issue and one that has traditionally plagued peer-to-peer networks. Designing a system that rewards storage provider nodes for remaining online for long periods of time should help combat these concerns. This challenge is recognized in the Storj paper when they say that the system must, “detect when a storage node stops storing data correctly or goes offline and then repair the data it had to new nodes.” Repairing data is done with erasure code reconstruction from remaining pieces. Erasure code reconstruction is something that is talked about in all of the papers here. Other systems help mitigate data loss as well by having multiple copies of any data spread out across the network in different locations. In some systems a Client seeking data storage can specify how many copies of their data they want to have stored in the network and this can not only provide redundancy but also higher availability. Potentially data could be downloaded from multiple sources at the same time similar to BitTorrent which could allow the system to offer much higher download speeds.

All data being stored in a decentralized network should be encrypted. This can be done client side or by the network itself and encrypting data before storing it on the cloud may be something that should be looked at even when using centralized cloud services. Ideally, encryption should be performed locally before data ever leaves a User's computer.

MaidSafe which was founded in 2006 in Scotland addresses their take on decentralized storage in many papers, one of which is, "The SAFE Network: a New, Decentralized Internet." (4) Unlike the other projects discussed here, MaidSafe actually predates Bitcoin which started in 2009. Like other projects, data is stored on the network in encrypted chunks. To address network availability, a minimum of four live copies of all data is stored at any one time. To address churn, nodes called data managers ensure that a new copy gets created every time a storage node goes offline. Content addressing is used in the Safe network and this allows the network to ensure that data on the network is not being over-stored. Furthermore, the creators of MaidSafe claim that their network is, "the first distributed hash storage system that supports deletion of data." (4) Additionally this can be done, "without explicitly listing the registered owners of chunks in the network."

In the paper, "IPFS - Content Addressed, Versioned, P2P File System", a system is theorized that would replace HTTP itself. They call HTTP, "the most successful distributed system of files ever deployed." IPFS takes inspiration from both peer to peer file sharing solutions like BItTorrent, and Git, the distributed source code version control system. It is claimed that Git's, "content addressed Merkle DAG model enables powerful file distribution strategies." Most notably, IPFS seeks to model all data as part of the same Merkle DAG.

IPFS was first developed by Juan Benet who is also one of the people who wrote the Filecoin paper some 4 years after this paper was published. The IPFS network is widely used today and represents the backbone that Filecoin is used on top of in order to incentivize the system. Actual data distribution in IPFS, according to this paper, happens by exchanging blocks with BitSwap which is a BitTorrent inspired protocol. BitSwap introduces a ledger like system to help rewards nodes in the network who share files and not nodes in the network who instead, freeload.

Other technological mainstays of IPFS, besides for version control (Git) and BitTorrent are distributed hash tables and SFS, or self-certified filesystems. Self-certified file systems (SFS) introduce both distributed trust chains and a shared global namespace. The name of an SFS filesystem can be verified by the public key offered by its server.

References

1. Vorick, David, and Luke Champine. “Sia: Simple decentralized storage.” *Nebulous Inc* (2014)
2. Storj Labs, Inc. “Storj: A Decentralized Cloud Storage Framework” (2018)
3. Benet, J., and N. Greco. “Filecoin: A decentralized storage network.” *Protoc. Labs* (2018).
4. Lambert, Nick and Benjamin Bollen. “The SAFE Network: a New, Decentralized Internet.” (2014).
5. Benet, Juan. “Ipfs-content addressed, versioned, p2p file system.” *arXiv preprint arXiv:1407.3561* (2014).