# Content Addressed P2P File System for the Web with Blockchain-Based Meta-Data Integrity

Chaitanya Rahalkar
*Department of Computer Engineering*
*Pune University*
Pune, India
chaitanyarahalkar4@gmail.com

Dhaval Gujar
*Department of Computer Engineering*
*Pune University*
Pune, India
dhvlgjr@gmail.com

***Abstract*—With the exponentially scaled World Wide Web, the standard HTTP protocol has started showing its limitations. With the increased amount of data duplication & accidental deletion of files on the Internet, the P2P file system called IPFS completely changes the way files are stored. IPFS is a file storage protocol allowing files to be stored on decentralized systems. In the HTTP client-server protocol, files are downloaded from a single source. With files stored on a decentralized network, IPFS allows packet retrieval from multiple sources, simultaneously saving considerable bandwidth. IPFS uses a content-addressed block storage model with content-addressed hyperlinks. Large amounts of data is addressable with IPFS with the immutable and permanent IPFS links with meta-data stored as Blockchain transactions. This timestamps and secures the data, instead of having to put it on the chain itself. Our paper proposes a model to use the decentralized file storage system of IPFS, and the integrity preservation properties of the Blockchain, to store and distribute data on the Web.**

***Index Terms*—IPFS, Blockchain, decentralized Systems, Peer-To-Peer Systems**

## I. Introduction

With the ever-expanding World Wide Web, the data generated on the web has grown vastly. The amount of data generated daily is at a staggering 2.5 quintillion bytes. [1] This pace is gaining constant momentum due to the inclusion of new IoT devices every day. Sensory data produced by IoT devices get bulkier as modern devices are added to the Internet. To counter the problem of data handling, many distributed file systems were introduced. The popular ones being Napster, BitTorrent, KaZaA, supporting millions of distributed users. Among all of them, HTTP - one of the oldest protocols on the Internet is the biggest distributed file system, when coupled with browsers allowing users to share files globally. With the increase in the scalability of the World Wide Web, the reliability of HTTP began to degrade. Keeping track of terabytes of data and moving these files over the Web is a difficult task. Several other protocols were introduced to tackle the problem of scalability and decentralization with the intention of replacing the well-established reign of HTTP. The other problem with HTTP is security and data integrity. As a countermeasure, the inclusion of Blockchain technology was introduced. Blockchain technology cannot be used to store the entire data due to its distributed ledger protocol. [3] This protocol states that every node in the Blockchain must preserve a copy of the data, on the chain. Hence, storing petabytes of data on the Blockchain is infeasible. This model proposes to store only the file metadata, summarizing necessary information about data, on the Blockchain. This data, being in bytes for a single file, reduces the overall size of the ledger. [8]

## II. Motivation

IPFS (InterPlanetary File System) is a proposed protocol that enhances HTTP. [2] We are entering the era of data distribution with new challenges like:

- Hosting petabytes of datasets
- Computing large data across organizations
- High volume, high definition, on-demand, real-time streaming of data
- Versioning and linking of massive datasets
- Preventing accidental disappearances of important files

To tackle all these problems, HTTP does not provide a scalable solution. Adding the middleware of Blockchain technology for preserving the file metadata helps maintain the integrity of the files that are stored. Blockchain technology induces its peculiar characteristics of data integrity, data security, and transparency to this file system. [7] Blockchain technology is a distributed ledger system that will preserve all the file metadata, including file size, author information, checksums, date of creation and modification, etc. To summarize, the distributed technology of IPFS and the data integrity feature of the Blockchain to preserve file-related information creates a full-fledged data serving model for the Internet. [6]

## III. History

The origin of the IPFS protocol dates back to the time when the DHT (Distributed Hash Table) was created. [5] The backbone of IPFS relies on the DHT protocol. It is a key-value store that uses distributed technology to store data. Key distribution takes place among nodes using a deterministic algorithm. Each node is assigned a portion of the hash table and it stores only the assigned data in the hash table. It uses advanced routing algorithms for data retrieval. The main disadvantage of DHT is data integrity and privacy. Since every node does not have a copy of all the data stored on the network, downtime of specific nodes may lead to data loss

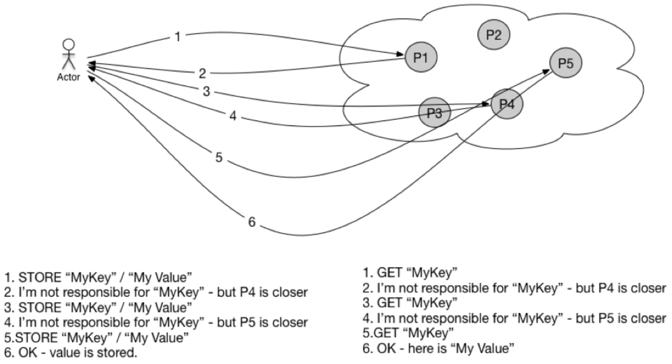
Actor
1
2
3
4
5
6
P1
P2
P3
P4
P5
1. STORE "MyKey" / "My Value"
2. I'm not responsible for "MyKey" - but P4 is closer
3. STORE "MyKey" / "My Value"
4. I'm not responsible for "MyKey" - but P5 is closer
5.STORE "MyKey" / "My Value"
6. OK - value is stored.
1. GET "MyKey"
2. I'm not responsible for "MyKey" - but P4 is closer
3. GET "MyKey"
4. I'm not responsible for "MyKey" - but P5 is closer
5.GET "MyKey"
6. OK - here is "My Value"

Fig. 1. Distributed Hash Tables

or non-availability of data. Also, the security of the data is compromised since data in the DHT nodes is not protected. Hence, Blockchain technology serves as an additional layer above DHT. In the Blockchain, copies of all the metadata of the files stored on the IPFS will be with every node. IPFS is a proposed replacement for the existing HTTP protocol.

## IV. Theoretical Concepts of Implementation

### A. Brief implementation

The IPFS is a distributed system similar to the BitTorrent protocol. Data is broken down into pieces and distributed across nodes in the network. The data element is assigned a unique IPFS hash that acts as an identifier for the file. The file is accessible via the IPFS hash. Every node in the IPFS network has its own IPFS daemon that communicates with other nodes in the network. When a file is uploaded to the network, a unique IPFS hash of the file is created and uploaded to the Blockchain network. Along with that, file-related metadata like author information, file size, and the file type is also uploaded.

Retrieval of files from the network is done via the network DHT. The network DHT uses advanced routing algorithms, beyond the scope of this paper, that find the data in log(n) time complexity, where n is the number of nodes in the network. After retrieving the data, the hash of the file is generated. The hash is searched on the Blockchain network, and if found, the metadata is retrieved from the network. This metadata is compared with the metadata of the file. Any mismatch indicates manipulation or file corruption without authorized permission. [10] The entire model is composed of four essential terminologies:

*1) Distributed Hash Tables:* A distributed hash table (DHT) is a type of a decentralized distributed system that works on the lookup mechanism similar to the hash table data structure. Key-value pairs are stored in a DHT, and nodes that are a part of the distributed system can efficiently retrieve the value associated with a given key. Keys are identifiers that map to particular values which in turn can be addresses, documents or arbitrary data. This allows a DHT to scale horizontally and handle continuous node arrivals, departures, and failures.

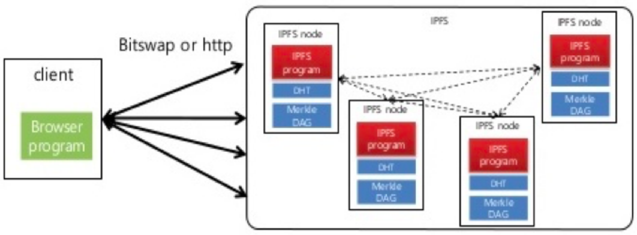

Fig. 2. Content Addressed Filesystem

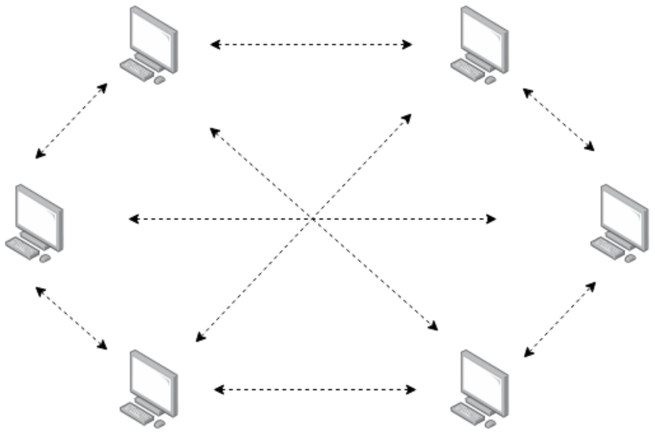

Fig. 3. Peer to Peer Network

*2) Content Addressed File System:* Everything on the World Wide Web is addressed with a URL that maps to the location of the file on the Internet. The IP address assigned to a website locates the file on the WWW. However, in a Content Addressed File System, the file is accessed based on its content, and not on its location. [9]

*3) Blockchain Technology:* Blockchain technology is a decentralized system similar to a ledger, a continuously-growing list of records without the possibility of tampering and revision. In a Blockchain, each node of the network stores the entire ledger data. So, the Blockchain mechanism differs completely from DHT, in which data is distributed among nodes. Every new transaction entering the Blockchain is validated using a process called mining.

*4) Peer-to-Peer Protocols:* Peer-to-Peer computing is a network in which each workstation has equivalent capabilities and responsibilities. They are typically situated physically near to each other and run similar networking protocols and software. Peers in a P2P network make a portion of their resources, available for other network participants, without the need for a central server. The P2P architecture is designed around the notion of equal peer nodes, i.e., peers are both suppliers and consumers of resources, simultaneously functioning as both ”clients” and ”servers” to the other nodes on the network.

Following are the components of the P2P Model:

1) Identities: Nodes are identified by a Node Id, the cryptographic hash of the node's public-key, generated with

S/Kademlia's crypto puzzle. Nodes store their public and private keys (encrypted with a passphrase). The Node Ids can be regenerated per daemon initialization.

2) Network: IPFS nodes communicate regularly with hundreds of other nodes in the network, potentially across the Internet

3) Routing: It is the mechanism to maintain information about location of specific peers and objects. The routing mechanism responds to both local and remote queries.

4) Exchange: It is a novel block exchange protocol also called - BitSwap, that governs efficient block distribution.

5) Objects: Every file in the P2P network is considered as a blob. This blob is made addressable with a Merkle DAG (Directed Acyclic Graph) of content-addressed immutable objects with links.

### *B. Need for Implementation*

The exponential growth rate of the World Wide Web has caused HTTP to start showing its limitations. There is a need to reinvent the protocol. Following are some of the limitations of HTTP:

1) HTTP is highly inefficient and costly: HTTP downloads a file from a single computer at a time, instead of getting segments from multiple computers simultaneously. With video delivery, a P2P approach has the ability to save 60% in bandwidth costs. IPFS allows the distribution of high volumes of data, effectively.

2) The web's centralization limits opportunities: The Internet has been accelerating innovation and has leveled the playing field. However, the increasing unification of control is a threat to this model.

3) Humanity's history is deleted daily: IPFS stores a versioned history of files and makes it simple to set up robust networks for mirroring of data.

4) Preservation of data integrity with Blockchain: Data tampering chances are incredibly high in the case of sensitive data. With the Blockchain middleware, data becomes immutable. Any attempt to change the metadata results in an invalid block, detecting the problem immediately.

### *C. Application Scope*

With the current demands of the industry, IPFS and Blockchain is a perfect pair to perform scalable and fault-tolerant tasks. Following are some of the use cases of this model:

1) Preservation of Massive Datasets: With the distributed technology, the model allows people to store large datasets, showing fast performance with a decentralized archiving system. Along with that, the integrity of the datasets can be preserved.

2) Sensitive Data Storage: Sensitive government documents, bond papers, contracts, etc. can be securely and safely stored with this model, avoiding cases of fraud. [12]

3) Content Delivery: Secured P2P content delivery saves millions in bandwidth, also providing better performance.

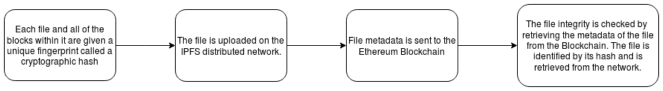
Each file and all of the blocks within it are given a unique fingerprint called a cryptographic hash
The file is uploaded on the IPFS distributed network.
File metadata is sent to the Ethereum Blockchain
The file integrity is checked by retrieving the metadata of the file from the Blockchain. The file is identified by its hash and is retrieved from the network.

Fig. 4. Model Flow

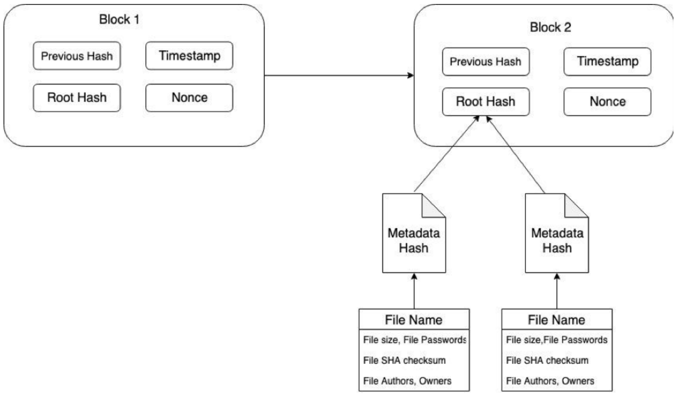

Fig. 5. Contents of Blockchain

## V. Model Flow

Figure 4 explains the flow of the model. When a file/folder is uploaded to the P2P network, a specific hash of the form "Qm..." is generated for the file and all the files within a folder. The files are uploaded on the P2P network, and the DHT is updated. These content addressed files can now be accessed via their unique hash. A proposed naming system called InterPlanetary Naming System (IPNS) analogous to the DNS is used to map file names to their unique hashes. The metadata of the file is then uploaded to the Blockchain. The Proof-of-Stake system of Ethereum [4] [13] is used for incentivizing. The authenticity and integrity of the file are verified by retrieving the hash and its data from the Blockchain.

Figure 5 shows the contents of the block in the Blockchain. Every block has a Merkle DAG(Directed Acyclic Graph) structure. The block comprises of four elements - The root hash, created from all the metadata hashes, the root hash of the previous block, the timestamp at which the block was mined, and the nonce. (used in the Proof-of-Work system)

TABLE I
File Metadata Stored on the Blockchain

| | File creation date | IPFS Hash Value | File Access Date | File Access Time | IPFS Hash of Modified File |
|---|---|---|---|---|---|
| 1 | 12/06/18 | $Qm....2$ | 14/06/18 | $12:45PM$ | $Qm....9$ |
| 2 | 16/06/18 | $Qm....19$ | 19/06/18 | $1:05PM$ | $Qm....25$ |

Table I shows the example file metadata. It comprises of parameters like file creation date, file creation time, access date, access time and other related file parameters.

## VI. Implementation Details

1) IPFS(Interplanetary File System) has a framework created for interfacing with the P2P network. This framework allows one to interact with the IPFS network. It connects with the IPFS daemon and translates and transfers requests via the TCP socket that it initiates.
2) An uploaded file is transferred to the IPFS network via the TCP socket. At the same time, file metadata is sent to the Ethereum Blockchain. A sample Blockchain entry is as shown in Table I
3) The IPFS gives a content-based address that can distinctively identify the file on the network.
4) When a file is to be retrieved from the network, the file integrity is verified by retrieving the metadata from the Blockchain. Each file is uniquely identified by a hash. The hash of the current file is verified with the one stored on the Blockchain. This validates the integrity of the file.

## VII. challenges

Despite the consistency of the IPFS protocol, a few issues are yet to be fully resolved. Firstly, the content addresses generated on IPNS are not human-friendly. These links can be shortened to easy-to-remember names using a Domain Name System (DNS), but this has the possibility of introducing an external point-of-failure for distribution of content. Many reports have suggested that IPNS can be slow at domain name resolution, with delays of up to a few seconds. Nodes may choose to "clear cached data" to save space since there is little to gain for the nodes by maintaining a backup for longer. Theoretically, this may lead to the disappearance of data if no nodes have a copy of that data. This is not a significant issue as of now, but for IPFS to be a viable, long-term solution, long term backups need to be strongly incentivized.

## VIII. conclusion

A secured and integrity compliant system was proposed using the P2P feature of IPFS and the tamper-proof principle of Blockchain technology. This model is a complete solution to the various problems faced by HTTP and data security. With minimal hardware requirements, any node in the decentralized network can serve data, improving bandwidth, latency, and availability. The four main components namely DHTs, Blockchain, P2P Networks, and Content Addressed File System, together, make the model a secured, reliable, and fault-tolerant system.

## References

[1] Marr, Bernard. "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read." Forbes, https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/. Accessed 5 Oct. 2019.

[2] Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. ArXiv, abs/1407.3561.

[3] Nakamoto, Satoshi. (2009). Bitcoin: A Peer-to-Peer Electronic Cash System.

[4] The Ethereum Wiki, 2019. GitHub, https://github.com/ethereum/wiki

[5] Labs, Protocol. IPFS Is the Distributed Web. IPFS, https://ipfs.io/. Accessed 6 Oct. 2019.

[6] Kelly, Mat, et al. "InterPlanetary Wayback: Peer-To-Peer Permanence of Web Archives." Research and Advanced Technology for Digital Libraries, edited by Norbert Fuhr et al., Springer International Publishing, 2016, pp. 411–16.

[7] Zyskind, G., et al. "Decentralizing Privacy: Using Blockchain to Protect Personal Data." 2015 IEEE Security and Privacy Workshops, 2015, pp. 180–84. IEEE Xplore, doi:10.1109/SPW.2015.27.

[8] Wang, Huaimin & Zheng, Zibin & Xie, Shaoan & Dai, Hong-Ning & Chen, Xiangping. (2018). Blockchain challenges and opportunities: a survey. International Journal of Web and Grid Services. 14. 352 - 375. 10.1504/IJWGS.2018.10016848.

[9] Benet, Juan. "IPFS - Content Addressed, Versioned, P2P File System." ArXiv:1407.3561 [Cs], July 2014. arXiv.org, http://arxiv.org/abs/1407.3561

[10] Chen, Y., et al. "An Improved P2P File System Scheme Based on IPFS and Blockchain." 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 2652–57. IEEE Xplore, doi:10.1109/BigData.2017.8258226.

[11] Anjum, A., et al. "Blockchain Standards for Compliance and Trust." IEEE Cloud Computing, vol. 4, no. 4, July 2017, pp. 84–90. IEEE Xplore, doi:10.1109/MCC.2017.3791019.

[12] Sarıtekin, R. A., et al. "Blockchain Based Secure Communication Application Proposal: Cryptouch." 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1–4. IEEE Xplore, doi:10.1109/ISDFS.2018.8355380.

[13] Wood, D.D. (2014). ETHEREUM: A SECURE decentralized GENERALISED TRANSACTION LEDGER.