# A Measurement Based Comparison of Centralized and Decentralized Storage Services

Akshay Raghavan
raghavan9@wisc.edu
University of Wisconsin-Madison

Dante Smith
dsmith67@wisc.edu
University of Wisconsin-Madison

## Abstract

The internet is undergoing a revolution - centralized services are being replaced by distributed and peer to peer alternatives. One such service that is becoming prominent today is the data storage network. Decentralized storage eliminates middlemen to provide cheaper and more private storing of data. This paper tries to analyze if there are actually merits to using or migrating to distributed storage networks such as BitTorrent, FileCoin and SafeNetwork over popular and proven cloud service providers like Dropbox, Google Drive and Microsoft OneDrive. Therefore we measure, analyze and compare properties such as performance, redundancy and cost between centralized and decentralized options to aid users in making informed decisions.

## 1 Introduction

### 1.1 Motivation

We are in an age of content creation and user content is becoming ubiquitous as mobile devices and storage services are becoming widespread and cheaper. Users are constantly looking for cheaper services with capabilities such as security and reliability to store their data. One popular option nowadays is personal cloud storage services provided by organizations like Google, Microsoft, Amazon etc. When users opt for these services, they store their personal files on the companies' servers and often forgo privacy concerns. Moreover, the big players in the cloud storage arena often dictate prices and customers are forced to comply.

There is an emergent market for decentralized storage that mitigates these issues. When a user file is uploaded to such a network, typically it is encrypted and broken up into multiple pieces often called chunks and stored in numerous nodes. Privacy is inherent as the file is encrypted and distributed to different providers. Any provider with spare space can sign up with such a network, providing an open market system where pricing is determined by demand and supply. This ensures the best possible price for the customers.

As we can see, the distributed P2P file storage seems to be an appealing alternative to personal cloud storage services but important questions need to be addressed as to its performance and reliability. Hence in this paper, we try to quantify these properties and compare the services provided by decentralized and centralized network storage providers. Most of the decentralized options include a blockchain and we also investigate the merits and drawbacks of using one as it relates to performance and security.

### 1.2 Research Problem

**To conduct a quantitative comparison between centralized and decentralized data storage networks with regards to performance, reliability, cost, and security.**

## 2 Proposed Approaches

We propose the following methods for achieving our objective.

### 2.1 Selecting Storage Options to Study

In order to start our study of comparing distributed storage options, we first choose several options in both the centralized and decentralized realms. This is done on the basis of popularity, underlying technology (for example blockchain-based), and functionality (such as permanent storage vs storage/retrieval/frequent updates). The selected services for centralized and decentralized providers are listed below:

- Centralized:
  - DropBox
  - Google Drive
  - Microsoft OneDrive
- Decentralized
  - BitTorrent
  - FileCoin
  - Safe Network

### 2.2 Selecting Appropriate Metrics for Study

Next, we try to identify the metrics that capture significant qualities and differences between the two domains. We measure various metrics to evaluate properties such as performance, reliability, security etc as listed below:

1. Performance
   a. Download time. This comprises of three different quantities:
      i. Time before file transfer (for example, time to lookup nodes with chunks)
         - There is a challenge to measure this for centralized services as it is difficult to monitor the activities on a proprietary server.
      ii. File transfer time
      iii. Reassembly time

   b. Upload Time. This comprises of two different quantities:
      i. Time before file transfer (for example, time to fragment file into pieces)
      ii. File transfer time
   c. Geographic location of file chunks. This metric helps us estimate the proximity of file chunks to the client and therefore the latency of file access.
   d. Number of file chunks. This will be useful to analyze the effects of chunking on redundancy and performance.
   e. Efficiency of storage. We identify how efficiently files are stored in providers by identifying the presence of file compression and deduplication.
2. Reliability and Fairness
   a. Redundancy
   b. Distribution algorithm for storing chunks
3. Cost per unit storage
   a. For example, how many dollars/GB data stored?
4. Privacy and Security
   a. What are the possible encryption schemes and protection methods employed by these services?

### 2.3 Selecting Techniques to Measure These Metrics

We propose estimating these metrics using active measurement techniques in real-time networks.

- Active measurements:
  This includes uploading and downloading files and using tools such as tcpdump on both the network nodes and the client to quantify the metrics
- Testnets:
  Decentralized networks typically provide test networks for deploying and validating beta versions. We use this capability in distributed p2p services like Filecoin to monitor the activity on not only the clients but also the storage network nodes.

## 3 Related Work

### 3.1 Storage management and caching in PAST, a large-scale, persistent peer-to-peer storage utility [1]

This paper analyzes a large-scale peer-to-peer persistent storage utility called PAST. The analysis focuses on the storage management and caching system. The authors also examine how PAST minimizes fetch distance and balances query load for popular files. We also are trying to measure these aspects of the storage systems we are analyzing, and will be using the work in this paper as a guiding methodology on how to test these aspects. This paper also gives useful overviews on concepts like security, replica diversion, storage imbalance, file encoding that we will be examining in this project.

### 3.2 Benchmarking Personal Cloud Storage [2]

This work highlights how to identify the capabilities of a cloud storage service such as data compression, chunking, deduplication, delta transmission, etc and establishes useful metrics like start-up synchronization time, completion time and protocol overhead for comparison of multiple providers. This work helped us identify what metrics to choose to effectively compare decentralized and centralized services.

### 3.3 BitTorrent (BTT) White Paper [3]

This work helps understand the functioning of BitTorrent, the classic peer to peer file transfer software. The paper elaborates on the tit-for-tat sharing protocol, how/why it evolved to incorporate a blockchain and the flaws in the current model. It gave us insights on how BitTorrent used swarms to group peers and trackers to lookup nodes with required file chunks.

### 3.4 A Digital Fountain Approach to Reliable Distribution of Bulk Data [4]

This paper describes a potential protocol for applications to reliably distribute bulk data to a large number of heterogeneous clients at times of their choosing (the authors call this protocol the “digital fountain”). In their description of the protocol, they also highlight aspects of networks that are important to evaluate to determine the functionality of the protocol. We chose to model the metrics we are tracking after the list provided in this project.

### 3.5 Filecoin: A Decentralized Storage Network [5]

This paper gives a technical overview of the protocol, network architecture and use cases. It also gives a formal definition of decentralized storage networks including its properties like fault tolerance, data integrity etc. It further outlines implementation of proof of replication, a proof-of-storage scheme used by FileCoin. It also talks about the retrieval and storage markets that govern the usage of the FileCoin network.

### 3.6 Online Data Backup: A Peer-Assisted Approach [6]

This paper studies how spare bandwidth and storage space of end-hosts complement that of an online storage service. Specifically, it looks at the data placement for the end hosts and the allocation of the spare bandwidth at those hosts. The paper then is able to show that using good data placement and bandwidth allocation policies can result in storage space from a cloud provider performing temporarily as well as a traditional client-server architecture. In our project, we are studying some solutions (for example, BitTorrent) that utilize bandwidth of remote end hosts, so this paper will give useful insight into how this might work.

### 3.7 The SAFE Network Primer [7]

This work details how the SAFE Network functions. This is an example of a decentralized storage option similar to those that we are studying in this project. It also gives a good overview of important concepts that we are reviewing, including how nodes and clients share information in a decentralized network, example security policies that these kinds of networks use, protocols used in the network, and an overview of the architecture (which can give insights to how architectures decentralized networks are generally setup).

### 3.8 Bitcoin: A Peer-to-Peer Electronic Cash System [8]

This paper was essential to learn how contemporary blockchains work, how distributed consensus is achieved using proof-of-work and transactions are validated in a trustless environment. Studying how Bitcoin's blockchain worked gave us a foundation to understand other blockchains (for example, FileCoin) as they are built with similar technologies.

## References

[1] Peter Druschel Anthony Rowstron. 2001. Storage management and caching in PAST, a large-scale, persistent peer-to-peer storage utility. *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles* (Oct. 2001), 188–201. https://doi.org/10.1145/502034.502053

[2] Marco Mellia Herman Slatman Idilio Drago, Enrico Bocchi and Aiko Pras. 2013. Benchmarking Personal Cloud Storage. *IMC '13: Proceedings of the 2013 conference on Internet measurement conference* (Oct. 2013), 205–212. https://doi.org/10.1145/2504730.2504762

[3] BitTorrent Inc. 2019. BitTorrent (BTT) White Paper. (Feb. 2019). https://www.bittorrent.com/btt/btt-docs/BitTorrent_(BTT)_White_Paper_v0.8.7_Feb_2019.pdf

[4] Michael Mitzenmacher John W. Byers, Michael Luby and Ashutosh Rege. 1998. A Digital Fountain Approach to Reliable Distribution of Bulk Data. *ACM SIGCOMM Computer Communication Review* 28, 4 (Oct. 1998), 56–67. https://doi.org/10.1145/285243.285258

[5] Protocol Labs. 2017. *Filecoin: A Decentralized Storage Network*. https://filecoin.io/filecoin.pdf

[6] Pietro Michardi Laszlo Toka, Matteo Dell'Amico. 2010. Online Data Backup: A Peer-Assisted Approach. *2010 IEEE Tenth International Conference on Peer-to-Peer Computing (P2P)* (Oct. 2010), 1–10. https://doi.org/10.1109/P2P.2010.5570003

[7] MaidSafe. 2021. The Safe Network Primer. (Nov. 2021). https://primer.safenetwork.org/

[8] Satoshi Nakamoto. 2009. *Bitcoin: A Peer-to-Peer Electronic Cash System*. https://bitcoin.org/bitcoin.pdf