1kx

Challenges in Decentralized Data

Storage and Databases

Jacob Eberhardt

@1kxnetwork
@Jacob_Eberhardt

1kx

Decentralized Storage

@1kxnetwork
@Jacob_Eberhardt

Storage

1kx

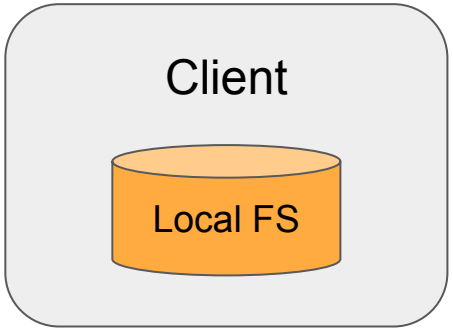
- Storage is defined as **the retention of retrievable data**
- This part is about **storing files**, i.e., uninterpreted BLOBs

Local Storage

1kx

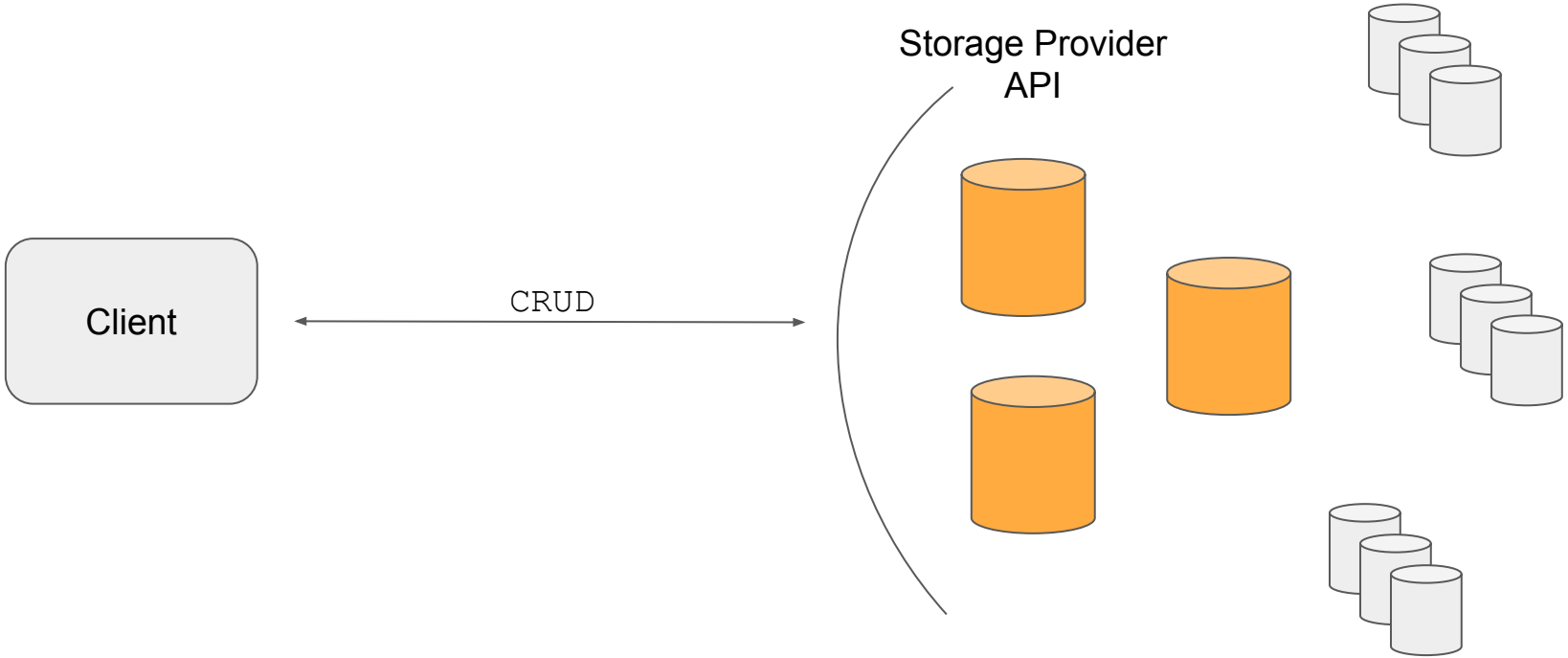
Trivial case: Client writes to local file system

- Single point of failure
- Bottleneck when other clients request data
- Client is responsible for security



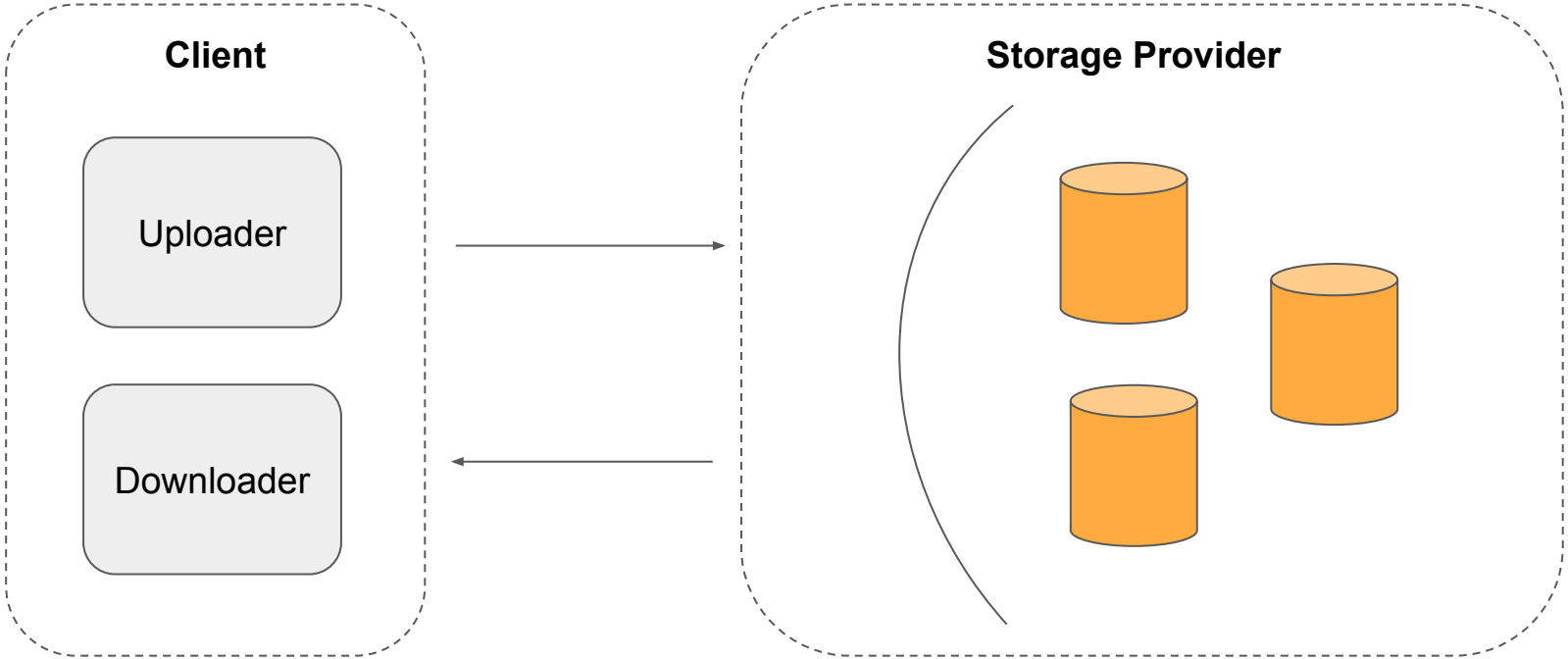
Centralized Storage: Architecture

1kx



Centralized Storage: Roles

1kx



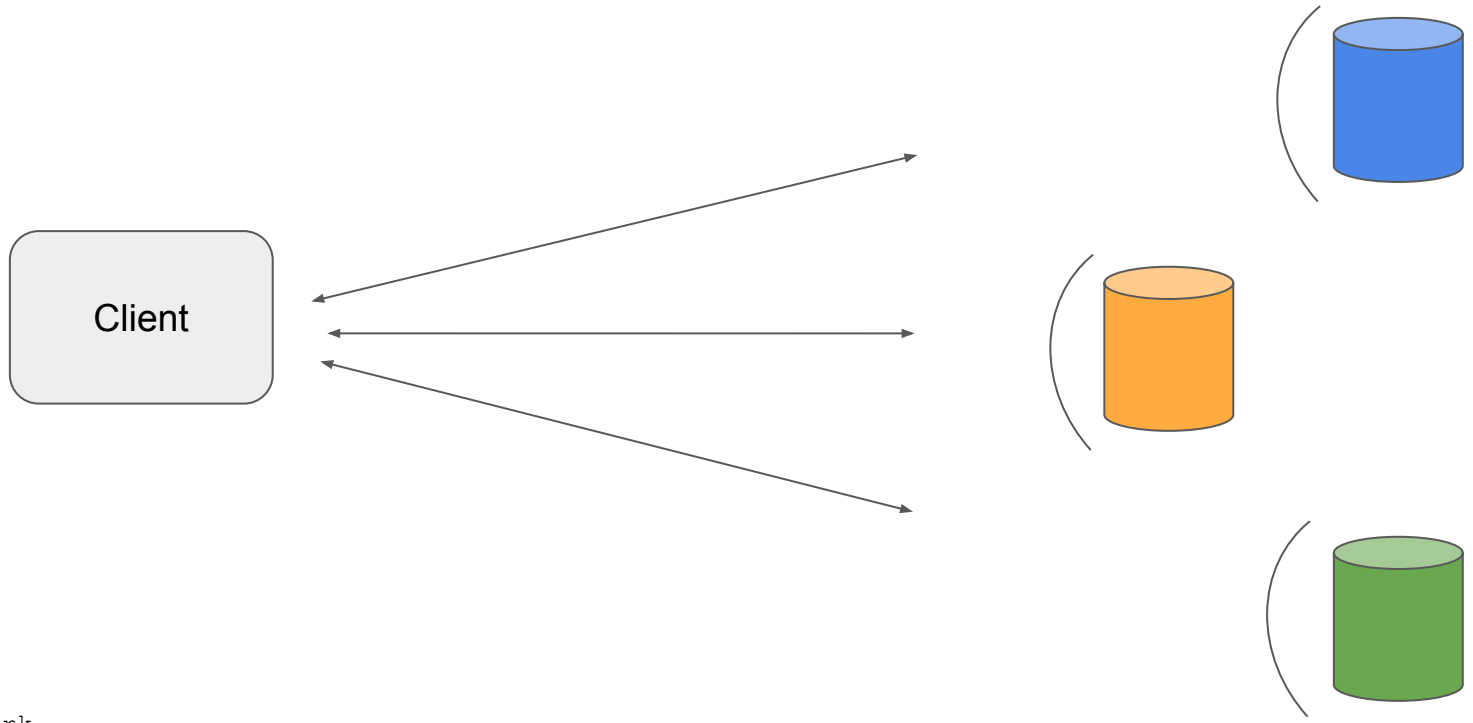
Amazon S3

1kx

- 11 9s durability (99,99999999999%)
- 4 9s availability (99,99%)
- 22 USD/TB/Month
- Client-Side and Server-Side Encryption
- Capacity scales ~linearly by adding more hardware
- Complex backend system, but very simple API
- Trusted required w.r. to availability, durability and security
- Economic incentive for Amazon to behave correctly

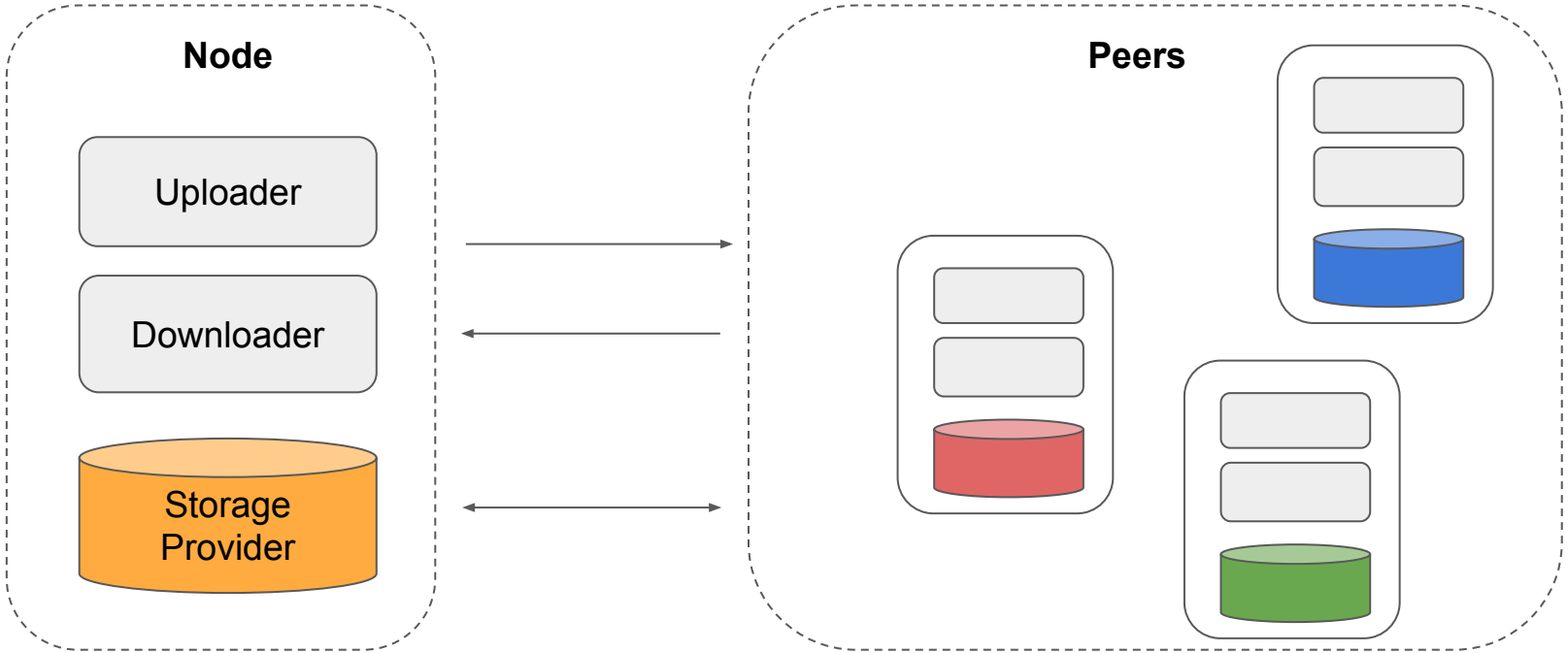
Decentralized Storage: Architecture

1kx



Centralized Storage: Roles

1kx



@1kxnetwork
@Jacob_Eberhardt

We have this!

1kx

- Bittorrent (2001)
  - Peer-to-peer filesharing network
  - Torrentfiles contain checksums and link to tracker nodes who forward to seed nodes

- IPFS (2015)
  - Content-addressable peer-to-peer decentralized file system
  - Users pin and cache files at their discretion

⇨ Non-incentivized decentralized storage

@1kxnetwork
@Jacob_Eberhardt

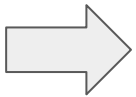
Non-incentivized Decentralized Storage

1kx

Fully decentralized file storage network

However

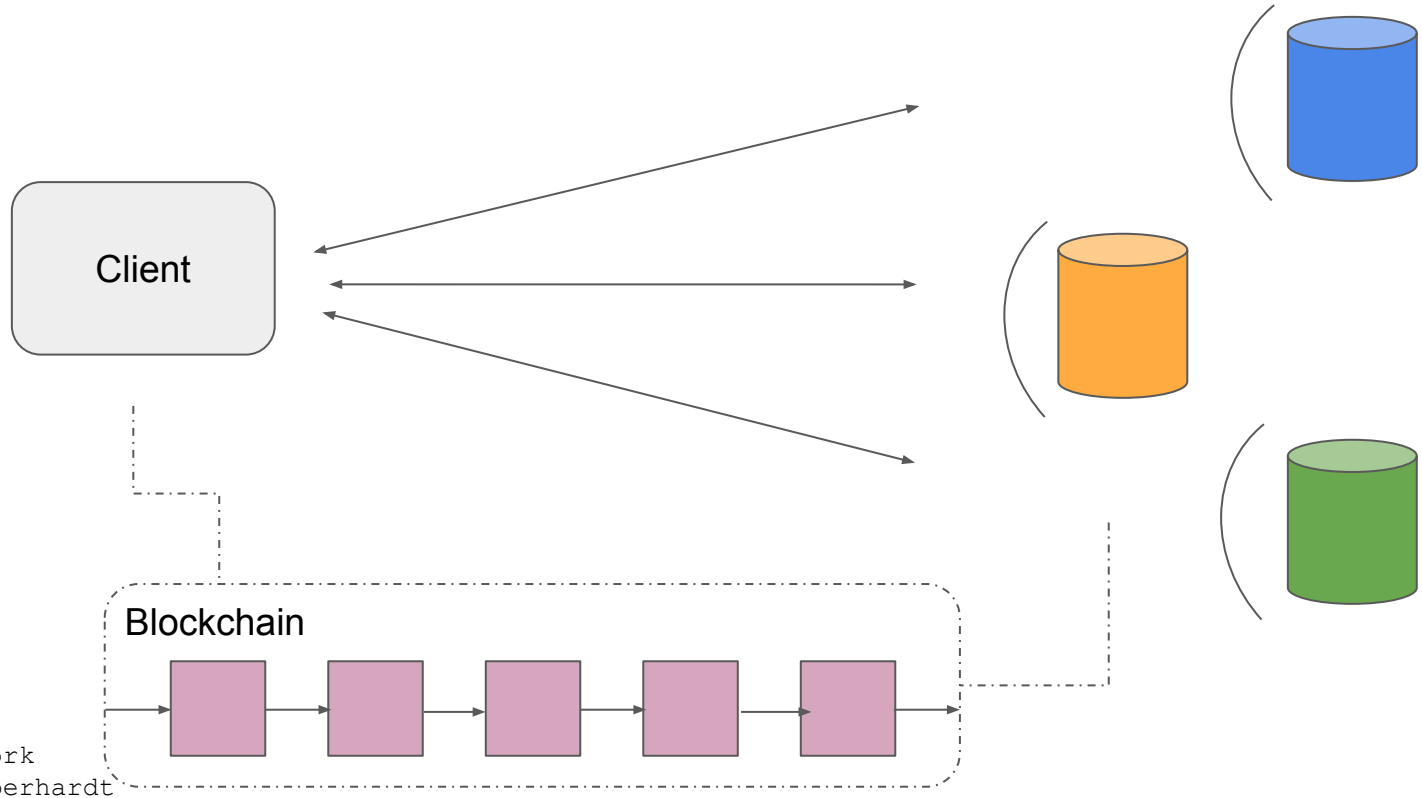
- Peers can shut down at any time
- Files can be lost
- Requests can be refused
- “Symmetric” participation pattern expected
  - Leeching peers, "freeloader problem"



No availability, durability, or performance guarantees

Incentivized Decentralized Storage

1kx



@1kxnetwork
@Jacob_Eberhardt

Incentivized Decentralized Storage Systems

1kx

- Use **cryptoeconomic protocols to ensure desirable properties** of the storage system
- Backed by blockchains to support these protocols

Two different goals when designing such an incentive system

- Decentralized File Storage Service for End-Users
  - Contracts between client and storage provider
  - Payment for certain storage time interval and SLAs
- Permanent File Archive
  - Protocol to ensure that no file is ever forgotten

@1kxnetwork
@Jacob_Eberhardt

Incentivized Decentralized Storage Challenges

1kx

The cryptoeconomic protocol needs to ensure

- Durability
- Availability
- Cost
- ...
-

Non-incentive system specific challenges

- Security
- Scalability
- Performance
- User Experience
- ...

@1kxnetwork
@Jacob_Eberhardt

Durability

1kx

Durability: Probability that data will survive permanently
“Will my data still be there in the future?”

Data must not be lost even in case of failures of storage providers!

Traditional Approaches:

- Replication
  - Store multiple copies of the data
- Erasure Coding
  - Transforms the original dataset into a larger dataset with n fragments such that the original data set can be recovered from k of the n fragments.

What is your approach to ensure durability?

@1kxnetwork
@Jacob_Eberhardt

Availability

1kx

Availability: Probability that a request is served successfully by the system when called.
"*Is my data accessible at the moment?*"

Data must be retrievable at any time, even if storage providers fail!

- How can a client be sure a file is available without having to retrieve it?
- How do you define SLAs
- Overhead of data-retrieval
  - Pre-payment before retrieval
  - Cryptoeconomic micropayment protocols

Are 4 9's achievable and desirable?

@1kxnetwork
@Jacob_Eberhardt

Attacks degrading Durability and Availability

1kx

- Sybil Attacks
  - Storage provider creates multiple identities
  - Gets paid to physically store multiple copies
  - Only stores data once
- Outsourcing Attacks
  - Storage provider claims to be storing data it does not physically store
  - Instead, it quickly retrieves data from other storage providers on request
- *Generation Attacks*
  - Storage provider claims to be storing a large amount of data
  - Instead, the provider efficiently generates the data on-demand using a small program
  - *This attack is only dangerous if "total amount of data stored" is used within the incentive layer*

How do you ensure independence of replica?

@1kxnetwork
@Jacob_Eberhardt

Incentive System Challenges

1kx

Crypto-economic protocol design is hard and a new discipline!

- How do you reliably detect malicious participants? E.g.,
  - Client requests data
  - Storage Provider serves data
  - Clients claims it was never served.
- Choice of Incentive Engine / Blockchain
  - Trust/Performance Tradeoff
- How do you ensure your protocol works correctly
  - Game theoretic proofs
  - Incentive-based argument

What's your approach to incentive design?

@1kxnetwork
@Jacob_Eberhardt

Cost

1kx

- Cryptoeconomic protocols cause overhead to ensure properties centralized providers get for free
  - Blockchain TX-fees
  - Coordination overhead
- Few centralized large providers.
  - Is the market competitive?
- Leverage existing hardware
  - There are end-users with unused disk capacity
  - The marginal cost of provisioning this capacity is low

How do prices compare today?

@1kxnetwork
@Jacob_Eberhardt

Security

1kx

- Large providers have security experts
  - How can decentralized storage providers be protected?
  - What are the main risks (data loss, data theft, DDoS)
- Key Management
  - Server-side encryption is no option
  - How can the challenges of key management be addressed
  - Are there recovery procedures?
- Encryption by default?
  - Storage providers cannot be trusted
  - How can encrypted data be shared?

How do you address peer protection and key management?

@1kxnetwork
@Jacob_Eberhardt

Scalability

1kx

- Centralized storage capacity scales linearly
  - What scalability behaviour can be expected for decentralized storage?
  - What are the bottlenecks (e.g., Blockchain)?

- Centralized storage can handle petabytes of data
  - How much data is stored in decentralized storage systems today?
  - What is the theoretical limit?
  - Which other properties degrade with size (e.g. latency)

What is your overall capacity today? What will it be in the future?

@1kxnetwork
@Jacob_Eberhardt

Performance

1kx

Latency

- Centralized providers have fast connections to internet backbones
  - What latency behaviour can we expect in decentralized systems?
  - Can we control the physical location of data storage to account for latency requirements?
- Cloud offerings allow co-location of application and data to reduce latency
  - Can we apply a similar approach in decentralized systems?

Throughput

- Centralized providers usually recombine data behind APIs before delivery
  - Can throughput in decentralized systems be higher due to parallel chunk retrieval and client-side recombination?

How do throughput and latency compare today and in the future?

@1kxnetwork
@Jacob_Eberhardt

User Experience

1kx

Centralized systems are usually called through an API with an API key

- What setup steps by the user are required before using the systems
  - Synchronize a blockchain?
  - Acquire a specific token?
  - Install a wallet just to store files?

- How can files stored be embedded
  - In websites?
  - In decentralized applications?

Which steps does a user need to go through to use the system?

@1kxnetwork
@Jacob_Eberhardt

Other Questions

1kx

- How tightly should blockchain and storage system be coupled?
  - E.g., Proof-of-Spacetime Mining in Filecoin
  - Use of Ethereum in Swarm
- Is there legal risk when participating?
  - What if a storage provider stores illegal files on a client's behalf?
  - Compliance with GDPR, e.g., right to be forgotten
- How are updates handled
  - Re-negotiation of terms with all participants?
  - Just store a new file? This is expensive!
- How can user privacy be protected in the P2P network?

@1kxnetwork
@Jacob_Eberhardt