Arweave Lightpaper

Version 0.9

Samuel Williams
William Jones

April 24, 2018

Contents

Abstract

Typical blockchains have several major well-known problems with data storage. These problems require new third-party protocols to be integrated on-top of existing blockchains, as fees are too high for on-chain storage to be feasible. Therefore, with typical blockchains there is always going to be a cost to access content, and content is never stored permanently. As the demand for data storage grows exponentially, the need for a decentralized low-cost data storage protocol that can scale is a necessity.

In this work we present Arweave – a new blockchain like structure called the blockweave. The blockweave is a platform designed to provide scalable on-chain storage in a cost-efficient manner for the very first time. As the amount of data stored in the system increases, the amount of hashing needed for consensus decreases, thus reducing the cost of

storing data. The protocol's existing REST API makes it trivially simple to build decentralised applications on top of the blockweave, reflecting Arweave's focus on the developer community and their ability to drive adoption of emerging and novel protocols.

In this paper, we also introduce novel concepts such as; block-shadowing, a flexibly-sized transaction block distribution algorithm that improves on current 'sharding' techniques by other blockchains, a self-optimising network topology, and a new consensus mechanism called proof of access.

1 Introduction

In this information age we often succumb to the illusion that because information is readily available, it can never be altered or lost. This is foundationally untrue [7]. While, in the internet, we have built a monumental system of decentralised information dissemination, we have yet to build a corresponding system of permanent knowledge storage. Modern history is full of examples of the destruction and loss of vital information, from fires at libraries and archives [9, 10, 3, 8], to book burning in authoritarian states [12, 11]. When we look up information on the internet, we are depending on being allowed access to centralised stores of that data. Access to the servers that hold this information can be revoked by their owners at any time. Similarly, as serving information on the internet requires the paying of server and upkeep costs, websites can often simply disappear when funds are no longer available. Further still, a number of governments are taking increasing steps to censor and remove access to politically sensitive information on the internet [13, 5, 4]. Equally with media and news organizations, where we once held physical and irrevocable copies, we now simply access the information and then discard it. It has become commonplace for media organisations to update the contents of their articles over time. While this provides a number of advantages over the previous system, most prominently, the ability to disseminate real-time updates about unfolding situations, it also allows important context to be lost or become obscured.

2 Background

All blockchain innovations sit on the shoulders of giants, including Bitcoin itself, a symphony of data structures, distributed networking and cryptography. We too have sought to further the space, solving specific shortcomings of existing blockchain networks, namely storage, and along the way a novel approach to transaction speeds. Most blockchain technologies today insist that a "full node" must maintain a copy of the entire blockchain in order to verify future transactions. While the Merkle data structures that make this possible are in and of themselves an impressive feat and add a layer of unparalleled security, we feel that some performance enhancements around this process could reduce the burden of synchronization for a full node. We present in section 4 several technologies that address block, node, and wallet

synchronization.

The full blockchain requirement is perhaps even more of a hindrance for existing blockchain technologies when it comes to storing data. In the case of Ethereum, a decentralised world computer, storage is incredibly costly using their native token. Arweave's primary motivation is to make permanent, immutable storage a reality, in the same way it is represented in Ethereum. However, high fees make this storage increasingly impractical. While it is possible to store data on the Ethereum, previous attempts have been impractical due to data storage costs.

Other blockchain technologies have focused on improving consensus algorithms between nodes, notably Stellar Lumens, and dPos architectures such as Ark and Neo. While this may improve transaction speeds, the burden of storage still remains the long term hurdle many of these networks will face. By focusing on solving storage first, we have experienced several performance enhancements that can be applied to facilitate high-throughput currency transactions.

3 Motivation

We have designed and implemented a blockchain network where permanent, low cost storage is a reality. Weaving storage access into consensus, combined with novel approaches to transaction bundling and arbitrarily sized blocks, creates a high-throughput cryptocurrency that improves on other cryptocurrencies like Bitcoin [10] and Ethereum [12]. In the past, archives (internet or otherwise) have typically been maintained by a single institution (or even individual), making them vulnerable to two primary forms of manipulation. The first of these is through the modification of documents during their storage [2]. The second is that the documents could have been forged or modified prior to their entry into storage [1]. For example, the many works attributed to Socrates that are believed to have been penned by his disciples [6]. Arweave solves both of these problems. Once the document is stored on the weave, it is cryptographically linked with every other block on the weave. This ensures that any attempt to change the contents of the document will be detected and rejected by the network. In this way, no subversion of the information on the weave is possible. Arweave is a browsable sister network to the internet, providing the long-term, permanent data storage features that the internet desperately needs but currently lacks.

A critical component of the Arweave system is designed for developers to easily build applications that interface with, create, and use data from the network. These apps, built with a language agnostic REST API, will act as a node in the network that listen to the network. The functions of these apps will be wide and varied, from decentralised and immutable social networks to discussion websites and news aggregators. In order to submit information to the weave, a small number of tokens will be required. These tokens will be used to pay miners for their work in maintaining the weave and network, as well as disincentivizing the propagation of spam. This

represents a great improvement over typical centralized storage systems. Similarly, it empowers individuals to ensure that the information they personally care about will be perpetuated into the future. The incentive to maintain the weave also increases as the network and documents will reinforce the value of the tokens. As these effects compound, we expect Arweave tokens to become a valuable asset for the information age; inseparably and intrinsically linked to a vast trove of important documents.

4 Technology

Arweave is built on four core technologies that work together to create low cost, high-throughput, permanent storage on a new blockchain. These innovations are:

- Blockweave
- Proof of Access
- Wildfire
- Blockshadows

While these technologies are intertwined, each plays a pivotal role in creating a new type of network suited for both fast transactions and low cost permanent storage.

4.1 Blockweave

A well known property of most blockchains is that every block must be stored to participate in validating transactions as a "full node". This is not the case with Arweave.

Instead, Arweave introduces two new concepts that allow nodes to fulfil key network functions without possessing the whole chain. The first of these concepts is the block hash list, a list of the hashes of all previous blocks. This allows old blocks to be verified, and potential new blocks evaluated effectively. The second of these concepts is the wallet list, a list of all active wallets in the system. This allows transactions to be verified without possessing the block in which the last transaction was used. Using these blockhash list and wallet lists synchronized by the network and available for download by the miners, nodes are able to join the network and participate in mining the weave almost immediately.

Further, instead of having each miner verify the entire block structure from the genesis block to the current block when they join the network, Arweave uses a system of 'ongoing verification'. When miners join the Arweave network, they will download the current block and retrieve the blockhash and wallet lists from the current block. Since these blockhash and wallet lists have been continuously verified through the ongoing progress of each block, new miners can start participating immediately without verifying the entire weave themselves. Full weave verification is of course available to any node that wishes to perform it. In this way, miners do not need to find the previous transaction associated with a wallet in order to verify a new transaction. Instead, miners would simply need to verify that the transaction has been appropriately signed by the wallet owners private key. To prevent recall block forging attacks, the hash of the blockhash list is

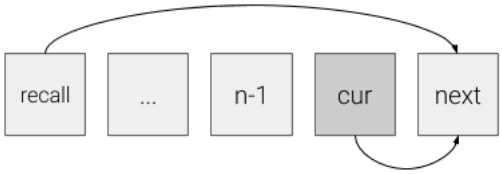

Figure 1: An illustration of the blockweave data structure, demonstrating the link to both the previous block and the recall block.

distributed with every new block.

4.2 Proof of Access

Arweaves consensus mechanism is based on proof of access (PoA) and proof of work (PoW). While typical PoW systems only depend on the previous block in order to generate each successive block, the PoA algorithm incorporates data from a randomly chosen previous block. Combined with the blockweave data structure, miners do not need to store all blocks (forming a blockchain), but rather can store any previous blocks, incentivised by PoA and wildfire, forming a weave of blocks, a blockweave. The 'recall block' to incorporate into the next block is chosen by taking the hash of the current block and calculating its modulus with respect to the current block height.

The transactions in the recall block are hashed alongside those found in the current block in order to generate the next block. When a miner finds an appropriate hash, they distribute the new block along with the recall block to other members of the network. This allows the other members of the network, even those without their own copy of the recall block, to independently verify that the new block is valid.

4.3 Wildfire

As a data storage system, Arweave requires not only the ability to store large amounts of information, but also to provide access to that data in the most expedient manner possible. Further, an important part of the Arweave system is costless access to data at the point of request. Subsequently, the Arweave has an added layer of incentives to encourage miners to share data freely.

Wildfire is a system that solves the problem of data sharing in a decentralised network by making the rapid fulfilment of data requests on the network a necessary part of participation. Wildfire works by creating a ranking system local to each node that determines how quickly new blocks and transactions are distributed to peers, based on how quickly they respond to requests and accept data from others. Peers are served in the order of their rank, with poorly performing peers being blacklisted from the network entirely. Peers are financially incentivised to stay well positioned in each other's rankings

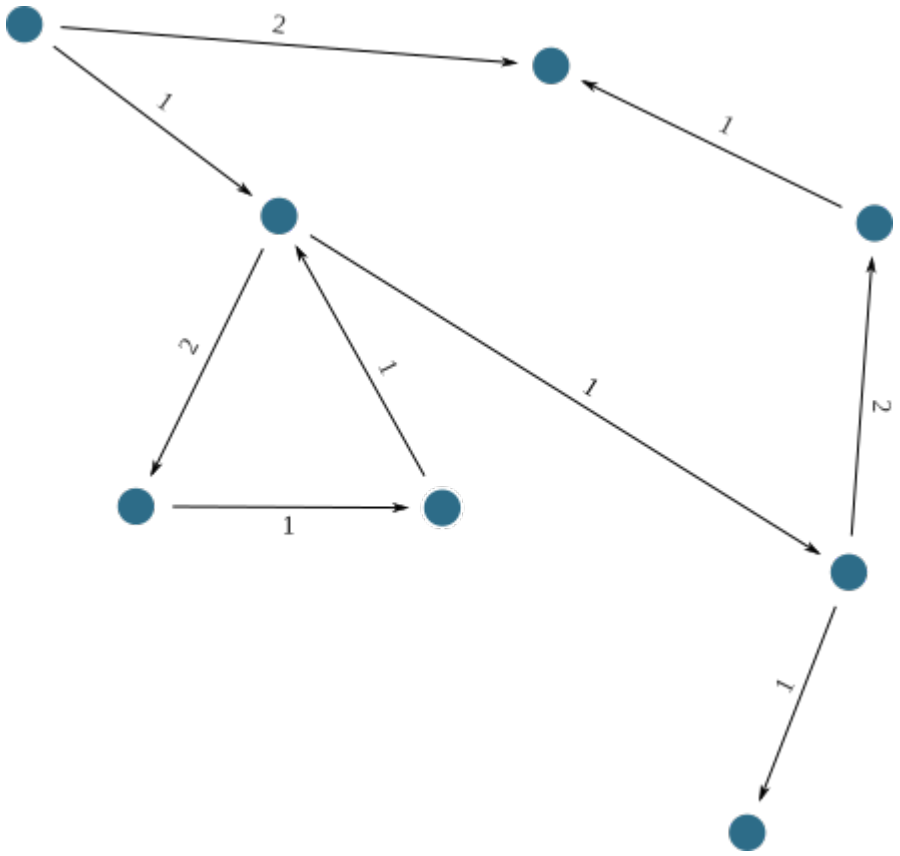

Figure 2: Illustration of the wildfire system. Each node ranks its peers based on how favourably these peers have behaved to them previously.

so that they can spend the largest amount of time efficiently mining.

This strongly encourages nodes in the system to behave in the most friendly manner possible to other peers, without cost to those who are receiving the data, even those who may potentially be making one-time requests. Even further, it creates a network topology that adapts to the most efficient routes for global distribution, as connections that allow fast transfer of new data around the system are prefered. In practise, the wildfire mechanism builds a network topology that maps the underlying physical connection substrate of the internet, adapting to changes in its architecture over time. Overall, the wildfire system ensures high speed distribution of new blocks and keeps data available with short latency.

4.4 Blockshadows

In a traditional blockchain system, when a new block is mined, each entire block is distributed to every node in the network, no matter how much of the block data a node already possesses. This is not only an enormous waste of data, but significantly slows down the rate at which a network can come to the consensus about a block. Arweave therefore introduces a new technology, blockshadows that not only minimises this waste of data, but enables fast block consensus and massive transaction throughput.

Blockshadowing works by partially decoupling transactions from blocks, and only sending between nodes a minimal block "shadow" that allows peers to reconstruct a full block, instead of transmitting the full block itself. These blockshadows specifically contain a hash of the wallet list and hash list, and in place of the transactions inside a block, only contain a list of transaction hashes. From this information (likely only a few kilobytes), a node who already holds all of the transactions inside the block and an up-to-date hash and wallet list can reconstruct an entire block of almost arbitrary size. To facilitate this, nodes will also immediately share transactions with one another, but only attempt to place transactions inside a block once they have a high certainty that other nodes in the network also have the transaction.

The result of this blockshadowing system is a fast and flexible block distribution system that allows transactions to be processed as fast as they can be distributed around the network, and consensus about blocks to be achieved at near network speed. Further, this system ensures transaction fees do not increase dramatically when network usage is high and a theoretical limit on transaction throughputs on an optimistic 100mbps network is around 5000 transactions per second.

4.5 Democratic Content Policy

To support the freedom of individual participants in the network to control what content they store, and to allow the network as a whole to democratically reject content that is widely reviled, the Arweave software provides a blacklisting system. Each node maintains an (optional) blacklist containing, for example, the hashes or substrings of certain data that it doesn't wish to ever store, and will never write to disk content that matches this. These blacklists can be built by individuals or collaboratively, or can be imported from other sources.

At a local level, these blacklists allow nodes to control their own content, but the sum of these local rejections also creates network wide content rejection. Content that is rejected by more than half the network will not only be rejected by each of those individual nodes, but will also be rejected by the wider network as a whole. This creates a democratic network-wide content rejection system that can merge blacklists across a variety of cultures and opinions into a tiny, specific blacklist of content that is universally reviled. This near universal, democratic blacklist shields the network from outside censorship by a small number of actors while still

allowing it the freedom to protect itself in a democratic manner.

4.6 Discussion

4.6.1 Storage Pools

One potential theoretical attack against the Arweave that has become extraordinarily large is that miners may work co-operatively to maintain a single copy of the weave, which they all access to retrieve recall blocks. While this kind of behaviour may at first seem problematic, this is not in fact the case. If such 'storage pools' were employed by a large proportion of the miners, the incentive for other miners to store rare blocks increases. This is because if the centralised stores become unavailable, miners with a copy of the rare blocks will be highly likely to receive the reward when that block becomes the recall block in the future. This self-interested behaviour provides a risk-offsetting function to the network, which scales as the potential for data loss (caused by centralised storage pools) grows.

5 Building Apps

Applications using the weave can be built using a simple REST API. The REST endpoints are HTTP and access the network directly, such that any Arweave wallet is capable of reading and writing data. The client only needs to bring their Arweave wallet to a website through a Chrome extension or native application with Arweave wallet integration, in order to read or write data from/to the network. There are several architectures that can be built on top of the weave.

5.1 Client-Server Architecture

Traditional web or native applications have a client-server architecture. A server running the cloud will be "Arweave enabled", interacting with one or more Arweave nodes, reading and writing data on behalf of clients. These services can be websites with clients as visitors, or they can be native applications passing client requests to a server operated by the developers. These servers will need to maintain a float of AR tokens in order to ensure that requests for writing data can be processed. Reading data from the weave however is still free using this architecture.

Monetization potential for this architecture is simple. A developer will need to accrue more value through advertising, monthly subscriptions or direct payments for a wrapper "credit" within their application, than the amount of AR tokens they are utilizing to power their storage. There are many applications for permanent immutable storage. For example, storing quantum resistant, encrypted legal case files, identity or medical records. While some legislation needs to accommodate sensitive information storage, geographical boundaries and the right to be forgotten, this can also be somewhat mitigated through encryption and key management. Several revenue generating models can be layered on top of the weave, with the primary value proposition being permanent immutable storage on-chain.

5.2 Serverless Architecture

Applications can live on the weave itself, accessed by a client through an Arweave enabled browser. Due to the ubiquity of browsers and proliferation of web technology, it makes most sense to store these applications as standard frontend web applications using HTML/CSS/JS. However, if the client's native application included an interpreter/parser for different languages such as LLVM bytecode or scripting language like Python, they could run on the client and perhaps benefit from the same upgradability found in web applications.

Developers will not only be able to deploy serverless applications to Arweave, these applications will also be able to write persistence and provable state to the network. Since Arweave does not impose a particular data structure, developers are free to store their data in the format that makes the most sense for them. If the application is best served by a highly optimized Merkle structure such as the one found in the Ethereum Virtual Machine (EVM), it can be easily implemented on the weave. If more text blob style storage is what the developer is looking for, this is trivial as well.

Serverless applications are extremely interesting as they can write their own data. Layering on distributed computation will, for example, allow the training of neural networks to store their results, possibly sharing their resultant models with other nets.

5.3 Event Based

In the early days of Twitter, there was a thriving ecosystem of cottage industry applications and developers building on top of the "firehost" APIs that were streaming tweets to anyone willing to pay for access. This is not the case anymore, and in the wake of the Facebook Cambridge Analytica fiasco, many "trusted partners" of these services that provided data analytics to their clients are being arbitrarily shut off.

Arweave is a decentralised network of public data and thus can never censor data access or the data itself, with the exception of democratically rejected content. This means that developers are free to build on top of Arweave and can listen for incoming data using the REST API. As events are triggered, the listeners will fire the appropriate function calls of the clients subscribed to those events. Developers need not fear being throttled or shut down, as the network is incentivised to provide them with reliable access to the data feed.

5.4 Trustless and Provable

Application architectures can be designed such that information needing to be stored and guaranteed as tamper-proof are easily implemented. Additionally, provably fair runtime code can be stored on the weave and interpreted directly by the client. Using the transaction ID of the content, the client can verify the payload from the weave prior to computation and be guaranteed that code they are running is both trustless and prov-

ably fair, i.e. it is the same code that other clients are running. This opens up interesting possibilities for trustless random number generators and other oracle-based services perhaps serving other blockchain networks.

6 Use Cases

Permanent storage has several use cases. Specifically, regulations requiring the archiving of documents up to a certain number of years. Provable media reporting, academic research and immutable records are becoming increasingly important in our modern world of echo chambers and proliferation of fake news.

6.1 Authenticity

Too often the legal system is tied up with litigation over the authenticity of documents. Arweave solves this problem by providing an indefinite and verifiable store of any digital content from an author. In 2017, the state of Delaware ruled to have blockchain evidence admissible in court proceedings. These records could dramatically speed up disputes over artistic attribution and intellectual property battles. The effects are twofold for the creative economy, allowing artists to license their work to others instantly and avoid frivolous litigation.

7 Conclusion

We have presented a new blockchain network powering low cost immutable data storage and a high-throughput cryptocurrency. The Arweave protocol is made possible through the use of a new blockchain-like data structure called the blockweave; flexible size transaction block distribution via blockshadowing; a new consensus mechanism reducing dependency on proof of work called proof of access; and a self-optimising network topology called wildfire. Much like the Bitcoin network, our technical advancements in isolation are not terribly complex; however, when combined to form the whole of the network, the emergent behavior is extremely powerful. We have seen from our testnet results that secure, reliable and immutable data storage is possible on a public, permissionless and decentralised network protocol. In addition to data storage, arbitrary size blocks make a secure high-throughput cryptocurrency possible without having to resort to complicated consensus mechanisms such as dBFT or dPoS.

Arweave is tightly woven into the fabric of the internet through its REST API and several revenue generating businesses are being built using the Arweave mainnet. Bridges between Arweave and other popular cryptocurrencies, secure computation, and smart contract protocols will enable a low cost and permanent data store to be easily integrated into the technology stack of decentralised applications. A fully globalized world of information and financial exchange requires permanent records. Through a combination of cryptography and distributed systems, we have provided the basis for those permanent recordings. We hope Arweave will become an essential companion to existing internet protocols such as the world wide web; working with

others to build a more open and transparent future.

References

[1] The national archives: Investigation into forged documents discovered amongst authentic public records. *http://discovery.nationalarchives.gov.uk/details/r/C16525*.

[2] North's ex-secretary tells of altering memos. *http://www.nytimes.com/1989/03/23/us/north-s-ex-secretary-tells-of-altering-memos.html*.

[3] The patent fire of 1836. *http://patent.laws.com/patent-act-of-1836/patent-act-of-1836-patent-fire-of-1836*.

[4] Mustafa Akgul and Melih Kirlidog. Internet censorship in turkey. *Internet Policy Review*, 4(2):1–22, 2015.

[5] Fernando Baez. *A universal history of the destruction of books: From ancient Sumer to modern Iraq*. Atlas Books, 2008.

[6] Anton-Hermann Chroust. Socrates–a source problem. *The New Scholasticism*, 19(1):48–72, 1945.

[7] Anne Frank and Storm Jameson. *Anne Frank's diary*. Vallentine, mitchell, 1971.

[8] Brewster Kahle. Fire update: Lost many cameras, 20 boxes. no one hurt., 2013. *https://blog.archive.org/2013/11/06/scanning-center-fire-please-help-rebuild/*.

[9] Birmingham Public Libraries. *Notes on the history of the Birmingham Public Libraries, 1861-1961*. Birmingham Public Libraries Birmingham, 1962.

[10] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008.

[11] Jonathan Rose. *The holocaust and the book: destruction and preservation*. Univ of Massachusetts Press, 2008.

[12] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 151, 2014.

[13] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. *Internet Censorship in China: Where Does the Filtering Occur?*, pages 133–142. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.