BIG DATA MEETS BLOCKCHAIN

DxChain

WHITEPAPER

www.dxchain.com

A Decentralised Big Data and Machine Learning Network Powered by a Computing-Centric Blockchain

by

The DxChain Team (support@dxchain.com)
Last Updated: June 20, 2018 Version 0.62 (draft)

Disclaimer: This White Paper is intended to be a technical overview. It is for community discussion purpose and not intended to be the final design.

NOTICE

PLEASE READ THE ENTIRETY OF THIS "NOTICE" SECTION CAREFULLY. NOTHING IN THIS WHITEPAPER CONSTITUTES LEGAL, FINANCIAL, BUSINESS OR TAX ADVICE AND YOU SHOULD CONSULT YOUR OWN LEGAL, FINANCIAL, TAX OR OTHER PROFESSIONAL ADVISOR(S) BEFORE ENGAGING IN ANY ACTIVITY IN CONNECTION HEREWITH. NEITHER DXCHAIN FOUNDATION LTD. (THE FOUNDATION), ANY OF THE PROJECT TEAM MEMBERS (THE DXCHAIN TEAM) WHO HAVE WORKED ON THE DXCHAIN NETWORK (AS DEFINED HEREIN) OR PROJECT TO DEVELOP THE DXCHAIN NETWORK IN ANY WAY WHATSOEVER, ANY DISTRIBUTOR/VENDOR OF DX (THE DISTRIBUTOR), NOR ANY SERVICE PROVIDER SHALL BE LIABLE FOR ANY KIND OF DIRECT OR INDIRECT DAMAGE OR LOSS WHATSOEVER WHICH YOU MAY SUFFER IN CONNECTION WITH ACCESSING THIS WHITEPAPER, THE WEBSITE AT HTTP://DXCHAIN.COM (THE WEBSITE) OR ANY OTHER WEBSITES OR MATERIALS PUBLISHED BY THE FOUNDATION.

All contributions will be applied towards the advancing, promoting the research, design and development of, and advocacy for a decentralised parallel-computing environment that supports big data and machine learning, as well as data storage, data exchange and big data computation for the data. The Foundation, the Distributor and their various affiliates would develop, manage and operate the DxChain Network.

This Whitepaper is intended for general informational purposes only and does not constitute a prospectus, an offer document, an offer of securities, a solicitation for investment, or any offer to sell any product, item or asset (whether digital or otherwise). The information herein below may not be exhaustive and does not imply any elements of a contractual relationship. There is no assurance as to the accuracy or completeness of such information and no representation, warranty or undertaking is or purported to be provided as to the accuracy or completeness of such information. Where this Whitepaper includes information that has been obtained from third party sources, the Foundation and/or the DxChain team have not independently verified the accuracy or completion of such information. Further, you acknowledge that circumstances may change and that this Whitepaper may become outdated as a result; and the Foundation is under no obligation to update or correct this document in connection therewith.

This Whitepaper does not constitute any offer by the Foundation, the Distributor or the DxChain team to sell any DX (as defined herein) nor shall it or any part of it nor the fact of its presentation form the basis of, or be relied upon in connection with, any

contract or investment decision. Nothing contained in this Whitepaper is or may be relied upon as a promise, representation or undertaking as to the future performance of the DxChain Network. The agreement between the Distributor and you, in relation to any sale and purchase of DX is to be governed by only the separate terms and conditions of such agreement.

By accessing this Whitepaper or any part thereof, you represent and warrant to the Foundation, its affiliates, and the DxChain team as follows:

(a) in any decision to purchase any DX, you have not relied on any statement set out in this Whitepaper;

(b) you will and shall at your own expense ensure compliance with all laws, regulatory requirements and restrictions applicable to you (as the case may be);

(c) you acknowledge, understand and agree that DX may have no value, there is no guarantee or representation of value or liquidity for DX, and DX is not for speculative investment;

(d) none of the Foundation, its affiliates, and/or the DxChain team members shall be responsible for or liable for the value of DX, the transferability and/or liquidity of DX and/or the availability of any market for DX through third parties or otherwise; and

(e) you acknowledge, understand and agree that you are not eligible to purchase any DX if you are a citizen, national, resident (tax or otherwise), domiciliary and/or green card holder of a geographic area or country (i) where it is likely that the sale of DX would be construed as the sale of a security (howsoever named) or investment product and/or (ii) in which access to or participation in the DX token sale or the DxChain Network is prohibited by applicable law, decree, regulation, treaty, or administrative act, and/or (including without limitation the United States of America, Canada, New Zealand, People's Republic of China).

Without prejudice to the generality of any term of any document between you and the Distributor or the Foundation, the Restricted countries shall be as follows: USA, China, New Zealand, Canada, Cuba, North Korea, Serbia, Tunisia, Somalia, Zimbabwe, Congo, South Sudan, Sudan (north), Sudan (Darfur), Turkey, Iran, Iraq, Libya, Syria, Ethiopia, Yemen, Sri Lanka, and Venezuela.

The Foundation, the Distributor and the DxChain team do not and do not purport to make, and hereby disclaims, all representations, warranties or undertaking to any entity or person (including without limitation warranties as to the accuracy, completeness, timeliness or reliability of the contents of this Whitepaper or any other materials published

by the Foundation). To the maximum extent permitted by law, the Foundation, the Distributor, their related entities and service providers shall not be liable for any indirect, special, incidental, consequential or other losses of any kind, in tort, contract or otherwise (including, without limitation, any liability arising from default or negligence on the part of any of them, or any loss of revenue, income or profits, and loss of use or data) arising from the use of this Whitepaper or any other materials published, or its contents (including without limitation any errors or omissions) or otherwise arising in connection with the same. Prospective purchasers of DX should carefully consider and evaluate all risks and uncertainties (including financial and legal risks and uncertainties) associated with the DX token sale, the Foundation, the Distributor and the DxChain team.

The information set out in this Whitepaper is for community discussion only and is not legally binding. No person is bound to enter into any contract or binding legal commitment in relation to the acquisition of DX, and no virtual currency or other form of payment is to be accepted on the basis of this Whitepaper. The agreement for sale and purchase of DX and/or continued holding of DX shall be governed by a separate set of Terms and Conditions or Token Purchase Agreement (as the case may be) setting out the terms of such purchase and/or continued holding of DX (the Terms and Conditions), which shall be separately provided to you or made available on the Website. In the event of any inconsistencies between the Terms and Conditions and this Whitepaper, the Terms and Conditions shall prevail.

This is only a conceptual whitepaper describing the future development goals for the DxChain Network to be developed. This Whitepaper may be amended or replaced from time to time. There are no obligations to update this Whitepaper or to provide recipients with access to any information beyond what is provided in this Whitepaper.

All statements contained in this Whitepaper, statements made in press releases or in any place accessible by the public and oral statements that may be made by the Foundation, the Distributor and/or the DxChain team may constitute forward-looking statements (including statements regarding intent, belief or current expectations with respect to market conditions, business strategy and plans, financial condition, specific provisions and risk management practices). You are cautioned not to place undue reliance on these forward-looking statements given that these statements involve known and unknown risks, uncertainties and other factors that may cause the actual future results to be materially different from that described by such forward-looking statements, and no independent third party has reviewed the reasonableness of any such statements or assumptions. These forward-looking statements are applicable only as of the date of this Whitepaper and the Foundation and the DxChain team expressly disclaims any responsibility (whether express or implied) to release any revisions to these forward-looking statements to reflect events

after such date.

The use of any company and/or platform names or trademarks herein (save for those which relate to the Foundation or its affiliates) does not imply any affiliation with, or endorsement by, any third party. References in this Whitepaper to specific companies and platforms are for illustrative purposes only.

This Whitepaper may be translated into a language other than English and in the event of conflict or ambiguity between the English language version and translated versions of this Whitepaper, the English language version shall prevail. You acknowledge that you have read and understood the English language version of this Whitepaper.

No part of this Whitepaper is to be copied, reproduced, distributed or disseminated in any way without the prior written consent of the Foundation.

Abstract

DxChain Network is a big data and machine learning network which is powered by a computing-centric blockchain with a native protocol token (also called "DX"). Its end users can potentially use this network as a data exchange platform to trade data and as a business intelligence analytics platform to analyse data for supporting business insights. Unlike Bitcoin and Ethereum, where the incentives are driving miners to sustain the huge amount of computation needed to maintain the blockchain consensus, the DxChain Network provides miners incentives based on the usefulness of the work that they passively provide: storage and computation. Since this storage platform is designed through using a principle of decentralisation, the protocol has a mechanism to control the reliability of the file access. In garnering the benefits of P2P network and Hadoop HDFS file systems, the robustness and accessibility is intended to be ensured. DxChain Network will incorporate Hadoop in its system as it is an industry-proven big data platform. DxChain Network is designed to assign several roles to manage and schedule jobs in the system which achieves the computation goal in a centralised parallel computing system. Built on top of the DxChain Network, a collection of tools would be developed to speed up the computation and analysis process. Additionally, machine learning algorithms can potentially be built upon these to facilitate more computing-driven tasks. At the time of retrieving files, DxChain Network is also designed to support computation which provides more flexibility of data exchange. With the smart contract, this platform is intended to be especially useful for decentralising data, building and running distributed applications.

This White Paper introduces the DxChain Network, a decentralised big data and machine learning network, powered by a computing-centric blockchain with four major innovations:

- A new decentralised computing framework to introduce Probable Data Computing and verification game;
- A new chains-on-chain design to orchestrate a master chain and two side chains—Data Side Chain and Computing Side Chain;
- Incorporate Hadoop to DxChain Network to facilitate big data and machine learning;
- Flexible and powerful DxChain Network-based system architecture dedicatedly designed to support most business data exchange and data analytics requirements.

Table of Contents

Chapter 1

Blockchain History and Motivation behind DxChain Network

Blockchain is an undeniably ingenious technological invention at the heart of Bitcoin [28] and other cryptocurrencies[13, 11, 23]. It is the brainchild of the person or group of people known by the pseudonym Satoshi Nakamoto.

A blockchain is a form of digitised and decentralised public ledger; it is used in all cryptocurrency transactions. The distributed ledger is a type of database that is shared, replicated and synchronised among the members of a network. It records the transactions, such as the exchange of assets or data, among the participants in the network.

Since the Bitcoin network started on January 2009, many other cryptocurrencies have emerged. The most popular consensus protocols are still based on the idea from Nakamoto either Proof of Work (PoW) [19] or Proof of Stake (PoS)[27, 35]. However, the consensus-based protocol by Nakamoto uses a huge amount of computational power to maintain the blockchain itself but does not provide useful work for the community.

Proof of Useful Work (uPoW)[9, 25] has been proposed to utilise this computational power. However, there are only a few cryptocurrencies that really use uPoW. Plasma[21] uses the MapReduce idea to distribute computational tasks from the main chain to its side chains; doing this could decrease transaction latency and increase process throughput. However, the main focus of Plasma is still on processing transactions.

Morpheo[26] is another platform that supports machine learning computations. It is made up of four different parts: 1) a local client running on data provider machines; 2) a cloud storage; 3) cloud computation resources and 4) a private blockchain network. This

system uses cloud computation resources to create a trustworthy environment; therefore, it protects the privacy of the data and avoids the verification of the correctness of the computation. However, this ultimately is a centralised system of computation.

Golem[9] is another platform that provides supercomputer services. Golem builds a transaction framework on top of Ethereum and the code is executed in the sandbox. The end users will publish the code into the community, in which validators that review the codes approve and rate the transaction model.

Unfortunately, there are no existing solutions in the market to provide a decentralised parallel-computing environment that supports big data and machine learning. DxChain Network incorporates Hadoop with blockchain to fill this gap. DxChain Network is envisaged as a comprehensive platform that provides data storage, data exchange and big data computation for the data in the system.

The next two chapters will provide an overview of the function and system design, while the following chapters will give more detailed explanations.

Chapter 2

Introduction to DxChain Network

Functionally, the entire system is designed as a platform to provide big data and machine learning related computation services with the support of decentralised data storage services. Database query and business intelligence tasks are performed for data that could not be owned by any commercial company but can be traded and stored on the DxChain Network. This platform is designed to connect personal computers, as well as specifically-designed miner machines, to potentially facilitate easier data computation at a lower cost.

DxChain Network provides two fundamental capabilities: computation and storage. Currently, only big companies have the capability to run big data tasks, because not only can these companies afford expensive hardware but these companies also own most of the existing consumer data. For example, Google possesses the navigation data of whoever that uses Google Maps. If Google Maps users could find a way to sell their navigation data, then they will do so for economic reasons. Users will not worry about their privacy since the data has already been seen and used by Google. If third parties can sell navigation data, combined with other commercial data, then advertising campaigns will be more accurate and customised. Most importantly, big companies like Google and Facebook will not monopolise the usage of these data. This is the dream of decentralisation.

The DxChain Network is designed to serve as a data trading platform for users who want to sell their data. Big data and machine learning computation is intended to make the use of data easier and more flexible.

Chapter 3

DxChain Network Design Principle & Architecture

The architecture of DxChain Network is adapted from the IPFS[11], Hadoop HDFS[22], GFS[31], FileCoin[12], IOTA[15], IoTeX[23], Plasma[21], TrueBit[20], morpheo[26] and Golem[9]. The entire system is designed based on the principle of incentivisation, which means that miners can maximise the utilisation of unused storage resources and facilitate decentralised big data computations with lower costs and much conveniences. The following section will describe the architecture of the DxChain Network. The chains-on-chain design, storage and computing with consensus protocol as well as the incentive mechanism will be explained in the subsequent sections.

3.1 DxChain Network architecture overview

DxChain Network provides decentralised big data and machine learning computation with the support of decentralised data storage. To achieve this complex project, the DxChain team is designing a specific chains-on-chain structure— which manages the master chain, storage chain and computation chain, to reach the consensus and to provide the incentive mechanism in a Byzantine environment.

A peer-to-peer storage network, such as InterPlanetary File System (IPFS)[11], Swarm and Storj, provides a flexible and extensible file system where the data-frame and schema can be build. The storage chain is built on top of the storage data model, in which Proof of Spacetime (PoSt)[24] validates whether the storage provider really stores the files.

DxChain Network will start with Hadoop, the industry-proven big data platform, as its computation engine. The Hadoop running elements, such as the job tracker, task tracker and workers, communicate with one another through using a computational chain to synchronise running jobs. With two consensus mechanisms: verification game and Provable Data Computation (PDC), this chain would incentivise nodes that provide computation power.

The master chain orchestrates the storage chain and the computation chain to maintain the master chain blocks, where the participating nodes may provide storage and computation power to obtain incentives.

Figure 3.1 shows the high-level DxChain Network infrastructure. From the figure, one can see how the master chain, storage chain and computation chain work together.

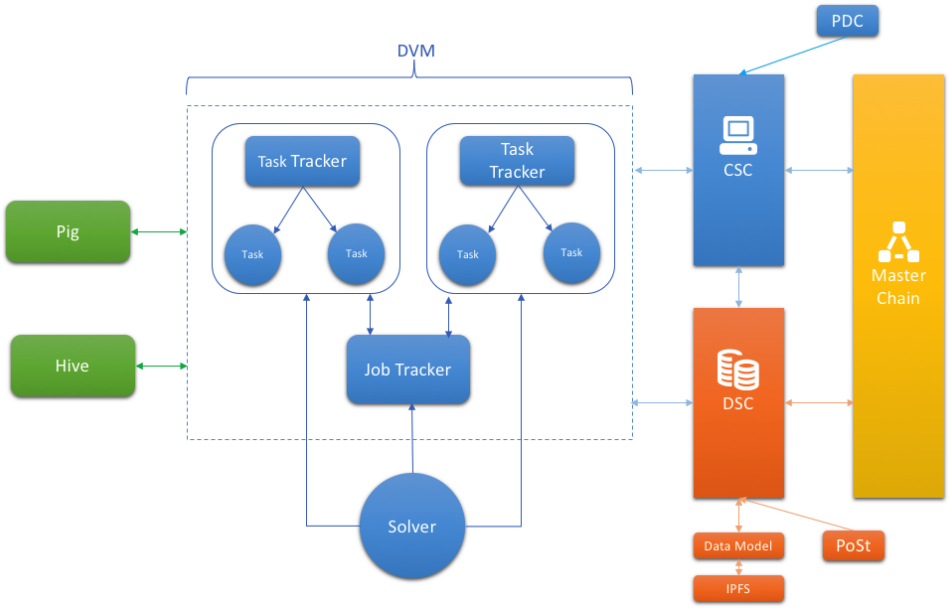

Figure 3.1: Illustration of High-level DxChain Network Infrastructure.

3.2 Chains-on-chain structure

The DxChain team is designing the three-chain structure with two unique properties: 1) master and side chains structure, 2) immutable master chain and elastic side chains.

Property 1 splits the chains according to their functions. Property 2 solves the problems of scalability, throughput and latency of the blockchain.

The master chain is in charge of maintaining transactions. Based on experience from Bitcoin and Ethereum, this master chain is kept immutable.

The side chains in the DxChain Network are in charge of storage and computation. These two functions are time-bounded, so that the blocks do not need to indefinitely store expired data. The DxChain team is designing a chain shortening algorithm which could remove unused data to save valuable block space.

3.3 Computing

Most of the popular blockchains are used for financial activity transactions, so their computations are simply checking transactional records on the chains; this type of computation does not require a lot of resources. However, the computations running on the DxChain Network are intended to support more general-purpose processing, such as database query and MapReduce work. Bitcoin uses blockchain to store all its transaction records to achieve consensus in the network and this consensus needs more than half of the active nodes to agree. It is impossible to store all computation states in a blockchain to ensure computation consensus; therefore, the DxChain team uses two mechanisms to ensure the correctness of the computations: verification game algorithm and Provable Data Computation (PDC).

Verification game algorithm is designed as a system with three main types of roles: solver, verifier and judge. This interactive system can testify to the correctness of a computation procedure without wasting too much computation power. Provable Data Computation (PDC) is a statistical framework to testify to the correctness of computation results with few copies of redundant computing.

Using verification game algorithm and PDC, DxChain Network is envisioned to solve the problems of outsourced computation and decentralised mining. Therefore, MapReduce work could potentially be performed on this platform.

3.4 Hadoop

Hadoop[1] is an industry-proven big data platform which includes HDFS for file storage, YARN or MapReduce for job scheduling, as well as a collection of tools to run distributed

jobs. Since Hadoop is a centralised system, it needs coordinators to schedule jobs. The DxChain Network is designed to adapt the Hadoop ecosystem to a decentralised environment.

In order to orchestrate the job tracker, task tracker and workers, a computation chain is designed to keep the computation states on the DxChain Network— it is used to validate the correctness of the computations and manage computation tasks. With the control of the computation states, a MapReduce pipeline can be implemented on the DxChain Network.

Built on MapReduce, a collection of tools such as Pig, Hive and Mahout is intended to be implemented for database operations and machine learning algorithms. Some business intelligence operations may be established through using these tools.

Chapter 4

DxChain Network: Blockchain Architecture

Bitcoin has been of great interest from among the decentralised cryptocurrency ledgers since 2009. However, at the same time, it is very difficult for the PoW consensus mechanism to be adapted to new demands and accommodate new innovation. In order to keep the benefits of blockchain, as well as make the chain light, fast and extendable, the DxChain team proposes a new design with two layers of blockchains chains-on-chain: master blockchain and side blockchains. The new system can more easily interoperate across chains for assets, data and messages. Besides, the side chains only carry the data storage and computing tasks, so technical innovation is not hindered.

4.1 Design overview

The master chain is similar to that of Bitcoin and Ethereum— it stores the ledger and asset information such as state, transaction and receipt, as well as the smart contracts. The master chain is good for storing small pieces of information since it is immutable. In order to support the complex data structure and computing information, the DxChain team proposes two side chains:

- Data Side Chain (DSC)— it is built upon a P2P distributed file storage system and stores the non-assets information.

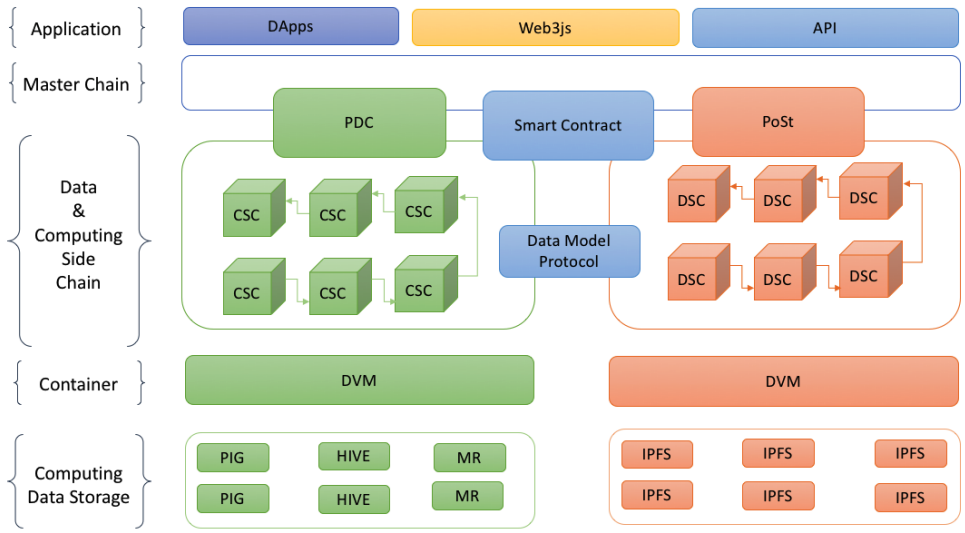

Figure 4.1: Illustration of DxChain Network Design Overview.

- Computing Side Chain (CSC)— unlike the hash mining (PoW) in Bitcoin, it is designed for useful work to solve real business problems and supports the specific computing task on the DxChain Virtual Machine (DVM).

The computing unit can read data from the DSC and also write the result back to the DSC. After one job finishes, the final state, related cost and corresponding incentives are stored on the master chain via smart contract. The intermediate states and task-level transaction information are kept in either the DSC or CSC. With the smart contract across the master and side chains, the entire system can keep a low cost on the master chain, as well as high efficiency computing and data storage on the side chains.

4.2 System architecture

In the DxChain Network, as shown in Figure 4.1, the master chain manages the overall transactions and the other two sides chains. The DSC and CSC communicate with the master chain through the smart contracts of DxChain Network. Besides, the DSC and the CSC can interoperate each other through the chains-on-chain microservices, which includes the data and messages.

The two side chains are designed to solve the efficiency issue of data storage and computing. The block structure is extended to store more data in the block and reduce the frequencies of the whole chain loading at the same time. Each side chain has its own consensus methods: the CSC uses Provable Data Computation (PDC) and verification game while the DSC uses Proof of Spacetime(PoSt).

The master chain and two side chains perform different functionalities. Even though these chains are interconnected through smart contracts and microservices, physically, the chains are still independent and isolated. The master chain is completely unaffected even when the side chains are broken. Any potential damage of each individual side chain is entirely confined to itself.

4.3 Master Blockchain

The master chain uses an Ethereum-compatible data structure which is composed of hash-linked blocks. A block is a collection of relevant pieces of meta-information, together with information corresponding to the comprised transactions, states and receipts. The blocks are connected through hash pointers.

Unlike the Bitcoin UXTO-based model, DxChain Network uses the accounts-based model to store the transaction and asset information, which includes account states, transactions across accounts and receipts. Similar to Ethereum, there are two types of accounts: regular account and contract account. This data is structured in a Merkle Patricia Tree data model and stored in all the node of the network.

Valid transactions are added into the master chain to enable the asset transfers between the master chain and side chains. The side chains use the same token from the master chain or each can define its own secondary token with the network-defined rate.

4.4 Data Side Chain

The Data Side Chain (DSC) is built on a peer-to-peer storage network, such as IPFS or Swarm. The chain itself works as an incentive layer; it is not being used for data storage. Proof of Spacetime (PoSt) is used as the consensus method for the microtransactions. It provides the foundation for the decentralised storage network. In this case, the advantages include faster setting times, lower transaction fees, faster transaction speed, higher privacy and the ability for transparency.

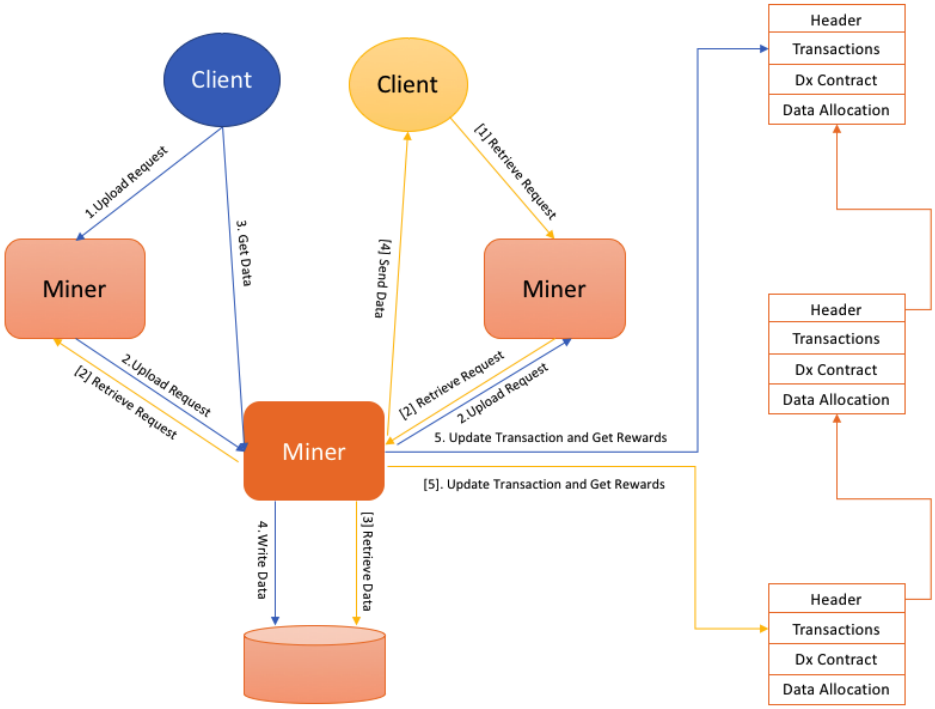

Figure 4.2: Illustration of Data Side Chain (DSC) Overview.

The data and files are broken down into many small pieces; they are stored into the peer-to-peer storage network, such as the InterPlanetary File System (IPFS). The meta-information and hash for each piece are stored in the chain, known as the file state, which similarly uses the Merkle Patricia Tree structure. Besides the hash value of each block and the file itself, the DxChain team has also designed a cross-chain URI for the file itself, so that the data can be easily accessible across the network and chains.

Between the data chain and P2P storage network, the DxChain team has designed a virtual logical layer, which includes the storage task giver, the miners for importing and exporting files and the verifier— this layer is termed as the DxChain Storage Layer.

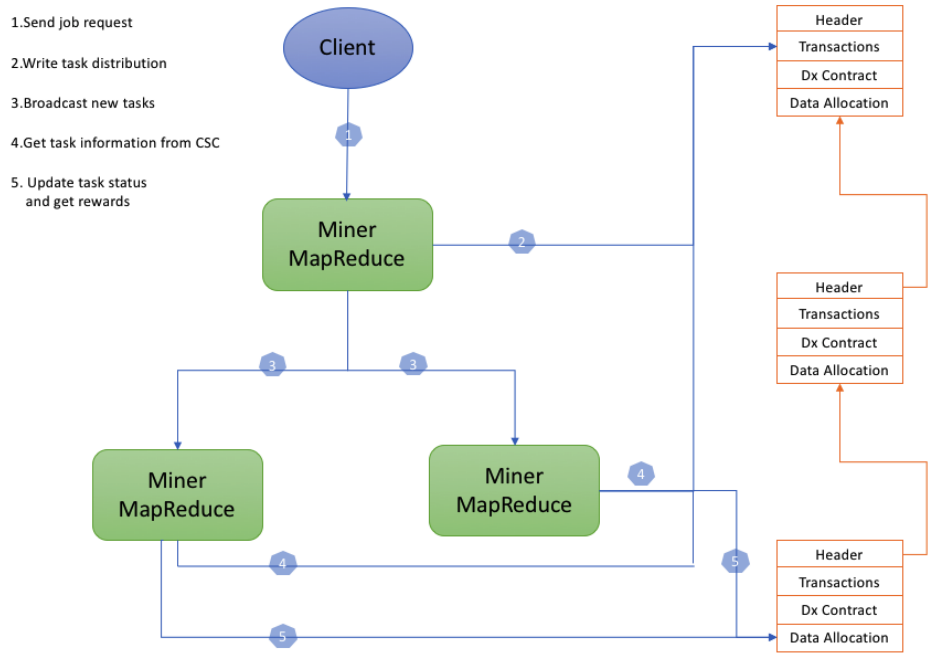

Figure 4.3: Illustration of Computing Side Chain (CSC) Overview.

4.5 Computing side chain

The Computing Side Chain (CSC) has a similar structure as the DSC — it is hash-linked and contains the header, transaction sets, contracts of the DxChain Network and data allocation. The transactions also use the Merkle Tree structure.

Figure 4.3 illustrates the details of how the CSC facilitates MapReduce operations. A client sends a computation request to the network and this request propagates through the network. Each miner uses the CSC to obtain tasks. When a task finishes, the working miner sends the confirmation to the CSC to update the task status and obtain incentives.

How the miners run MapReduce jobs will be explained in the later chapter.

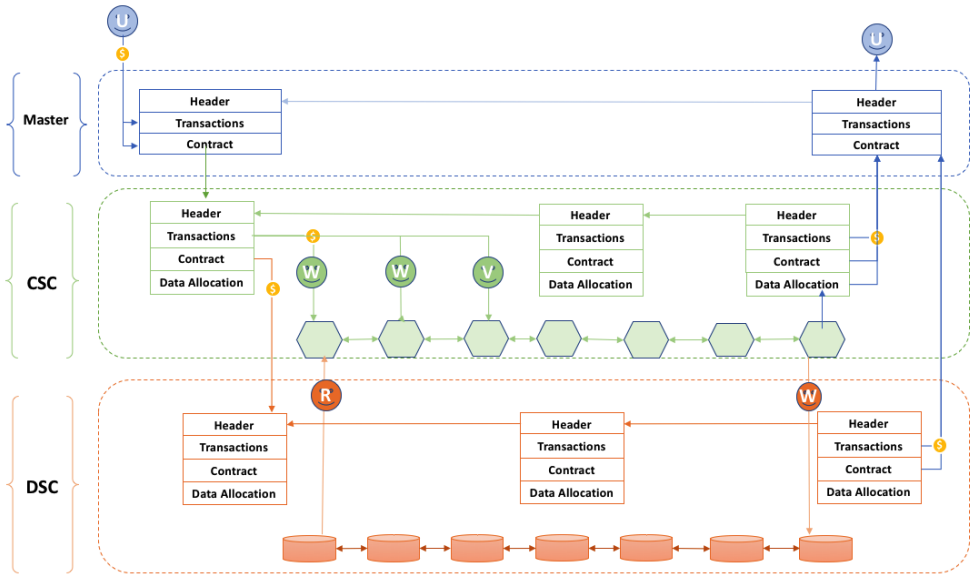

Figure 4.4: Illustration of DxChain Cross Chain Communication Protocol.

4.6 Communication cross chains

Figure 4.4 illustrates the DxChain Cross Chain Communication Protocol, showing an overview of the transactions across the whole network.

With a transaction flow across the whole system between the master chain, DSC and CSC, the illustration is a review of the communication protocol at a general level.

1. The task giver ("U") submits a task in the master chain. The miner in the master chain needs to check the following:
   - the block in valid format, including states, transactions, receivers, and contracts;
   - the orders are valid;
   - the pieces of proof are valid;
   - the cross-chain contract is valid;
   - the assets in the master chain are locked and is transferred to the CSC.
2. The solver and verifier in the CSC will load the code and data into the DxChain Virtual Machine (DVM) and execute the code inside the DVM including parallel computation and verification tasks. The miner in the CSC needs to:

- check the block format;
- check the credit and deposit are valid;
- check the task related-data and code are valid;
- verify the task result if need;
- read/write data from DSC if need;
- summarise the translations and transfer them back to the master chain.

3. The main job for the miner in the DSC is to provide the fundamental data storage and data transfer across the P2P network. The data writer-miner will store the data in the network and the data reader-miner will retrieve the data. The miner in the DSC needs to:

   - check the block format;
   - conduct maintenance on the health of the data as well as ensure its availability, integrity and security;
   - check that the credit and deposit are valid; and
   - transfer to the CSC if the asset in the master chain is locked.

4. After all the computing tasks, data storage and transfer tasks are completed, the overall transaction and asset will be transferred back to the master chain.

Chapter 5

Storage & Data Model

DxChain Network is designed as a decentralised computation network with the support of a decentralised storage network where files are stored for computation results and all kinds of intermediate computation states. The storage miners are intended to obtain incentives based on their continuous contributions to the storage chain. Apparently, the consensus-based protocols by Nakamoto, including PoW[19] and PoS[27, 35], do not fit this requirement. Proof of Spacetime (PoSt)[24] is a good choice to validate the provision of storage. It has been discussed in Chapter Four that the DSC is designed to manage storage tasks and that DSC will also be connected to the master chain to obtain incentives for the storage miner as well as the CSC for storing computation states. For the design of the DSC, please refer to Section 4.4. Next, the PoSt protocol and the data model will be explained.

5.1 Consensus protocol

Since Proof of Spacetime (PoSt) evolved from many of its predecessors, this section will start from Provable Data Possession and then move to PoSt.

5.1.1 Provable Data Possession

Provable Data Possession (PDP)[18] was introduced for allowing a client that has stored data on an untrusted server to verify that the server stored the original data without retrieving it. This model provided the first provably-secure scheme for remote data checking.

Definition 1. (Provable Data Possession Scheme (PDP)) A PDP scheme is a collection of four polynomial-time algorithms $(KeyGen, TagBlock, GenProof, CheckProof)$ such that:

$KeyGen(1^k) \rightarrow (pk, sk)$ is a probabilistic key generation algorithm that is run by the client to setup the scheme. It takes a security parameter k as input, and returns a pair of matching public and secret keys (pk, sk).

$TagBlock(pk, sk, m) \rightarrow T_m$ is an algorithm run by the client to generate the verification meta-data. It takes as input a public key pk, a secret key sk and a file block m and return the verification meta-data T_m.

$GenProof(pk, F, chal, \sum) \rightarrow V$ is run by the server in order to generate a proof of possession. It takes as inputs a public key pk, an ordered collection F of blocks, a challenge $chal$, and an ordered collection $\sum$ which is the verification meta-data corresponding to the blocks in F. It returns a proof of possession V for the blocks in F that are determined by the challenge $chal$.

$CheckProof(pk, sk, chal, V) \rightarrow \{`success', `fail'\}$ is run by the client in order to validate a proof of possession. It takes an inputs a public key pk, a secret key sk, a challenge chal and a proof of possession V. It returns where V is correct proof of possession for the blocks determined by $chal$.

A PDP system can be developed from a PDP scheme in two phases, Setup and Challenge.

- Setup: The client C is in possession of the file F and run $(pk, sk) \leftarrow KeyGen(1^k)$, followed by $T_m \leftarrow TabBlock(pk, sk, m)$. C stores the pair (sk, pk). Then C sends pk, F and $\sum$(T$_1$, ..., T_m) to S for storage.
- Challenge: C generates a challenge chal and an indicator of which blocks C want to prove S possess. C sends challenge and indicator to S. S runs

$$V \leftarrow GenProof(pk, F, chal, \sum)$$

  and send V back to C.
- Finally, C can check the validity of the proof V via running $CheckProof(pk, sk, chal, V)$.

PDP scheme provides a solution of verifying the server S storing the File F at the time of the client C sending the challenge $chal$. In order to prove server S storing F for a period of time T, the client C must keep sending challenges to S to verify this in order to have a high probability of certainty of S storing F in a continuous time T.

5.1.2 Proof of Spacetime

PDP as described in Section 5.1.1 is a perfect fit for cloud computing storage but it has several limitations for a decentralised network, such as privately verifiable computing. The Proof of Spacetime (PoSt) scheme is designed to be an improvement over the PDP scheme and is fit a decentralised environment.

Threat models.

($SybilAttack$) An adversary A has Sybil identities $P_0, P_1, \ldots, P_n$, and makes each commit to storing a replica of F. The attack succeeds if $P_0, \ldots, P_n$ store less than n copies of F and produce n valid proofs that convinces the verifier V that F is stored as n independent copies.

($OutsourcingAttack$) Upon receiving challenge $chal$ from verifier V, an adversary A fetches the corresponding F from another storage provider P^* and produces the proof that A stored F.

($GenerationAttack$) If an adversary A is in a position of determine F, A may choose F such that A can re-generate F and produces the proof that A stored F.

Definition 2. ($Proofofspacetime(PoSt)$) A PoSt scheme is a collection of three algorithms $(Setup, GenProof, CheckProof)$ such that:

$Setup(1^k, F) -> (R^F, S_p, S_v)$ where S_p and S_v are scheme specific variable for P and V that depend on the File F and on a security parameter k.

$GenProof(S_p, R^F, chal) \rightarrow \pi^{chal}$ where chal is a challenge, R^F is a replica of F. $GenProof$ is run by the prover to produce a proof π^{chal} for a verifier V.

$CheckProof(Sv, chal, \pi^{chal}) -> \{`success', `fail'\}$ which check if a proof is correct. $CheckProof$ is run by V and convinces V if P has stored R^F.

A PoSt system must be complete and secure. A system is complete if any honest prover P that stores a replica of F can always produce valid proofs and convinces a verifier V. If a PoSt system is secure if this system must prevent the attacks in the attack model.

Prevention of Sybil attack. To ensure the independent physical storage of n copies of F could be handled by treating n different $F's$. For each file F, we could enforce the prover P encode F: $R^F_{ek} = Encode(F, ek)$ where each ek is unique for each replica of F. For

$$ek_i! = ek_j, R^F_{eki}! = R^F_{ekj}.$$

Instead checking the storage of F, we could check R^F_{ek}. Since ek is unique for each replica of F, no more than one sybil identities could provide the same R^F_{ek}. This encoding process must be reversible that means $F = Decode(R^F_{ek}, ek)$.

Prevention of outsourcing attack and generation attack. A time-bounded algorithm could prevent these two attacks. At the stage of $GenProof$, the prover P must return the proof within a time bound, otherwise P will be failed at the $CheckProof$.

$$T^{honest} = RTT^{v \rightarrow p \rightarrow v} + T(GenProof(Sp, R^F, chal))$$
$$T^{adversary} = RTT^{v \rightarrow p \rightarrow v} + T(GenProof(Sp, Encode(F, ek), chal))$$

PoSt chooses an encoding algorithm such as $T^{Encode} >> T^{honest}$. Choice of Encoding. In order to satisfy the condition $T^{Encode} >> T^{honest}$, we will choose a encoding function with the following properties: 1) it must be reversible, 2) the output should be determined from an encoding key ek, 3) information should not be compressible across replicas, 4) encoding running time should be scale with a tunable parameter. A Pseudo Random Permutation (PRP) can be slowed down using a block cipher in cipher block chaining mode. We will use it for the encoding f unction.

Zero-knowledge proof and public verifiable.

Publicly Verifiable Computing. The PoSt scheme requires publicly verifiable computing because decentralisation requires that every participating party could be able to verify the validity of the transactions.

In a zero-knowledge proof protocol, the prover proves a statement to the verifier without revealing anything about the statement other than that it is true, which protects the prover against any malicious verifier which attempts to gain more knowledge than what is intended. The protocol can be either interactive or non-interactive. The key difference with non-interactive proofs is that all interactions consist of a single message sent by the prover to the verifier. The following notation is used

$$NIZKPoK(\alpha, \beta) : a = g^\alpha \wedge b = g^\beta$$

to denote a non-interactive zero-knowledge proof of knowledge of the values and such that

$$a = g^\alpha \text{ and } b = g^\beta.$$

All values which are not enclosed in parenthesis are assumed to be known to the verifier. When one uses a non-interactive zero-knowledge proof to authenticate auxiliary data, the

resulting scheme is referred to as signature of knowledge [14]. Basically, a signature of knowledge scheme means that one in possession of a solution w to the problem x has signed the message m. For the above $NIZKPoK$, the following notation

$$SoK[m](\alpha, \beta) : a = g^{\alpha} \wedge b = g^{\beta}$$

is used to denote a signature of knowledge on message m.

Proof chain. The PDP scheme could only prove that a File F is stored at a given time. To ensure that the file F has been stored for a continuous amount of time, the verifier must send challenges periodically. It has a higher probability for continuous storage if the verifier sends challenges with higher frequency. However, proof chain provides another mechanism with a stronger assurance of continuous storage.

A proof chain is a verifiable data structure that chains a sequence of challenges and proofs. For iteration n, let $chal_n$ be the proof of iteration $n-1$. Give a randomness r and a collision resistant hash (CRH) function H, a proof chain C is constructed as follows:

1. $chal_0 \leftarrow H(r)$;
2. $chal_n \leftarrow H(n, V_{n-1})$,
3. $V_n = GenProof(chal_n)$,
4. $C_0 \leftarrow (chal_0, V_0), C_n < -(chal_n, V_n)$,
5. $\{`success', `fail'\} \leftarrow VerifyChain(C_0) = CheckProof(chal_0, V_0)$ and
6. $\{`success', `fail'\} \leftarrow VerifyChain(C_n) = VerifyChain(C_{n-1}) + CheckProof(chal_n, V_n)$.

A proof chain could provide the proof of continuous time storage without frequently asking for verification and the prover could prove the storage even when it is offline.

5.2 Data Model

The DxChain Network is designed to not only provide the data storage function; it also gives the flexibility of retrieving files at a more granular scale. This section will start with a discussion on storing standard file formats on the DxChain Network, for example, text files or binary file types.

Structured text data. More specialised forms of text files are structured formats such as CSV, XML and JSON. These types of formats can be presented using DxChain Network. With the schema of the data, DxChain Network provides the filter and search functionality supporting most of the data exchange requirements.

Binary data. Although text is typically the most common source data format stored, binary files such as videos and images are also supported. However, DxChain Network does not support the finer granularity search for binary data.

DxChain Network supports both schema-on-read and schema-on-write operations. The schema-on-read allows for the rapid landing of large amounts of data into the storage but requires extensive tagging of such data to ensure that it is generally usable across the network. The schema-on-write data storage requires a lot more up-front preparation and ongoing transformation of the incoming data, so it is more expensive to set up and maintain but has the advantage of storing the data in a more standardised and consistent fashion.

The schema-on-write is a mechanism supporting privacy preservation. In the following file, the columns "Name" and "SSN" should be encrypted based on the schema provided at the time of loading files to the network. At the time of retrieving files according to the descriptions, the schema gives the DxChain Network a flexible way to operate.

Table 5.1: Illustration of how data model works.

Name	SSN	GENDER	DESCRIPTION
James Bond	123-456-6789	M	Spy
Larry Page	234-567-7890	M	CEO
...	...	...	...

Name	SSN	GENDER	DESCRIPTION
🔒	🔒	M	Spy
🔒	🔒	M	CEO
...	...	...	...

5.3 Usage and sale of token

Usage of token. The native digital cryptographically-secured utility token of the DxChain Network (DX) is a major component of the ecosystem on the DxChain Network, and is designed to be used solely as the primary token on the network. DX will initially be issued by the Distributor as ERC-20 standard compliant digital tokens on the Ethereum blockchain, and these will be migrated to tokens on the native blockchain when the same is eventually launched.

DX is a non-refundable functional utility token which will be used as the unit of exchange between participants on the DxChain Network. The goal of introducing DX is to provide a convenient and secure mode of payment and settlement between participants who interact within the ecosystem on the DxChain Network. DX does not in any way represent any shareholding, participation, right, title, or interest in the Foundation, its affiliates, or any other company, enterprise or undertaking, nor will DX entitle token holders to any promise of fees, dividends, revenue, profits or investment returns, and are not intended to constitute securities in Singapore or any relevant jurisdiction. DX may only be utilised on the DxChain Network, and ownership of DX carries no rights, express or implied, other than the right to use DX as a means to enable usage of and interaction with the DxChain Network.

DX is required as virtual crypto "fuel" for using certain designed functions on the DxChain Network, providing the economic incentives which will be consumed to encourage participants to contribute and maintain the ecosystem on the DxChain Network. Computational and storage resources are required for running various applications and executing transactions on the DxChain Network, as well as the validation and verification of additional blocks / information on the blockchain, thus providers of these services / resources would require payment for the consumption of these resources (i.e. "mining" on the DxChain Network) to maintain network integrity; accordingly DX will be used as the unit of exchange to quantify and pay the costs of the consumed computational resources. DX is an integral and indispensable part of the DxChain Network, because without DX, there would be no incentive for users to expend resources to participate in activities or provide services for the benefit of the entire ecosystem on the DxChain Network. Users of the DxChain Network and/or holders of DX which did not actively participate will not receive any DX incentives.

In particular, you understand and accept that DX:

(a) is non-refundable and cannot be exchanged for cash (or its equivalent value in any other virtual currency) or any payment obligation by the Foundation or any affiliate;

(b) does not represent or confer on the token holder any right of any form with respect to the Foundation (or any of its affiliates) or its revenues or assets, including without limitation any right to receive future dividends, revenue, shares, ownership right or stake, share or security, any voting, distribution, redemption, liquidation, proprietary (including all forms of intellectual property), or other financial or legal rights or equivalent rights, or intellectual property rights or any other form of participation in or relating to the DxChain Network, the Foundation, the Distributor and/or their service providers;

(c) is not intended to represent any rights under a contract for differences or under any other contract the purpose or pretended purpose of which is to secure a profit or avoid a loss;

(d) is not intended to be a representation of money (including electronic money), security, commodity, bond, debt instrument or any other kind of financial instrument or investment;

(e) is not a loan to the Foundation or any of its affiliates, is not intended to represent a debt owed by the Foundation or any of its affiliates, and there is no expectation of profit; and

(f) does not provide the token holder with any ownership or other interest in the Foundation or any of its affiliates.

Sale of DX tokens. The Distributor of DX shall be an affiliate of the Foundation. The contributions in the token sale will be held by the Distributor (or its affiliate) after the token sale, and contributors will have no economic or legal right over or beneficial interest in these contributions or the assets of that entity after the token sale. To the extent a secondary market or exchange for trading DX does develop, it would be run and operated wholly independently of the Foundation, the Distributor, the sale of DX and the DxChain Network. Neither the Foundation nor the Distributor will create such secondary markets nor will either entity act as an exchange for DX.

Chapter 6

Computing Machine

The key of DxChain Network is to provide a decentralised computation environment. In a Byzantine environment, one has to have a validation mechanism to verify the correctness of the computation in order to outsource computation and to outsource against the adversaries who ask for incentives without executing the tasks. Verification game provides a framework to validate the correctness of the computation procedure and Provable Data Computation (PDC) provides a statistical scheme to find a corrected answer from a set of untrusted nodes with a small probability of being attacked. For the design of the computation chain CSC, please check Section 4.5.

The following will explain the verification game framework and then move on to Provable Data Computation scheme.

6.1 Consensus protocol of computation

Verification game and PDC use two different philosophies to design the validation process; DxChain Network uses both of them in the system for the different types of computation.

6.1.1 Verification game

In theory, verification game provides a framework which enables smart contracts with no need for trust to securely perform any computation task. Moreover, when compared with the computations in traditional Ethereum smart contracts, verification game vastly reduces the number of redundant network node computations.

In a verification game, there are several roles playing different functions in the system. The core roles are Solver, Challenger and Judges. A Solver is a miner who offers a solution to a given task and a Challenger is one who disagrees with the solution from the Solver. The Judges, who always give the correct computations, use extremely limited computation bandwidth.

The trick in a verification game to eliminate the computation of Judges narrows down the portion of the computation in dispute with a series of rounds. A matrix multiplication was used to illustrate the process. The resulting item with coordinate (i, j) is the dot product of the ith row in the first matrix and jth column of the second matrix. The dot product could be split into n steps if the vector length is n. The dot product is then broken down into more atomic computations of summation of scalar multiplication. In the dispute period, the Solver and Challenger will argue with each other to identify which scalar multiplication results in the disagreement. The Judges only need to check the scalar multiplication as compared with the whole matrix multiplication.

Verification game does not trust or rely on the reputation of its participants or any trusted party in the system. A deposit is needed to perform a task from both the Solver and the Challenger. For any faulty players, they will lose the deposit. This penalty mechanism will potentially eliminate the untrusted players with the passing of time. Since the Judges do not have a lot of computation to do, they are the entire community of Ethereum miners who reach verdicts through the Nakamoto consensus.

6.1.2 Provable Data Computation

There is not a single deterministic mechanism that exists which can prove the correctness of any computation. Since the DxChain Network needs some nodes to perform some computational work which should be validated, the DxChain team proposes a statistical way to solve this problem.

Setting. Let one consider an untrusted network with total online nodes N, there are M adversary nodes within this network.

Solution. A computational task could be broadcasted through the network. U nodes perform the task; the answer which was the first identical one that W nodes generated is chosen as the valid answer. Figure 5.1 shows how to determine the correct computation results.

Proof. If the wrong answer is chosen, that means that all the W nodes are adversary

nodes. The probability of this even happening is

$$P(\text{attack}) = \binom{M}{W} / \binom{N}{W}.$$

If N is high and M is 20% of N, $P(\text{attack}) < \epsilon$ with a very small W.

Table 6.1: Working principal of PDC.

W = 7 Computation: 1 + 2		
IDENTICAL RESULT	RESULT	NODE
1	3	N8
2	3	N9
1	1	N108
1	5	N523
3	3	N667
2	1	N928
3	1	N5627
4	3	N7763
5	3	N5326
$return(3, N8, N9, N667, N7763, N5326, N4, N3)$		

Empirical study.

Networks with active users of 100, 500 and 1000 nodes are chosen for this study. Among those active users, there exists 20, 100 and 200 adversary users respectively. Different $W's$ are tried to determine the threshold of the consensus protocol. When W equals 7, the $P(\text{attack})$ is at the level of 10^{-7} which is a very strong supporting statement. In Figure 6.1, the green line represents $(N, M) = (100, 20)$, the red line represents $(N, M) = (500, 100)$ and the blue line represents $(N, M) = (1000, 200)$.

6.2 Usage in DxChain Network

Both verification game and PDC have advantages and disadvantages. Verification game has a strong statement. However, it is an interactive verification process and the computation

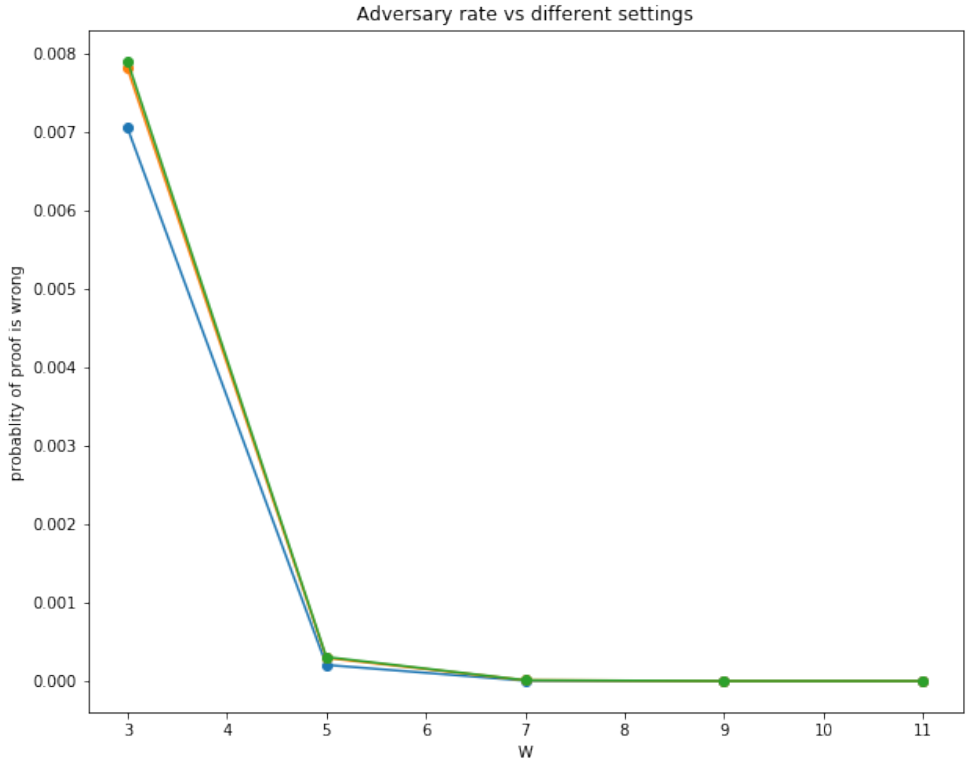

Figure 6.1: Experical study of PDC.

break-down is not easy to implement. PDC has a weak conclusion but it is enough for most distributed computing. PDC is very easy to apply to distributed computing.

DxChain Network uses both verification game and PDC for the different scenarios.

Chapter 7

Computing Platform

Nakamoto-consensus-based cryptocurrency uses lots of computational power to maintain the blockchain rather than providing real useful work for the community. Lots of Proof of Useful Work schemes have been proposed, from the computation for protein folding to Proof of Storage, however very few of them have a real application.

Although TrueBit says that its system is good for all computation, the technological details are still focused on financial transactions. Plasma uses the concept of MapReduce to scale up the computation for transactions but it is not the MapReduce computation that most people are talking about, which is distributed computing. DxChain Network introduces Hadoop, which is the system used as the industry standard for distributed computing platforms on blockchain.

7.1 History of Hadoop

Hadoop was developed based on three research publications: The Google File System[31], MapReduce: Simplified Data Processing on Large Clusters[16] and Bigtable: A Distributed Storage System for Structured Data[17] from 2003 to 2006. The design of Hadoop is using commodity hardware to build clusters for parallel computing for big data. Starting with the Apache Nutch project and moving to Apache Hadoop in 2006, it is fully commercialised and well used in the business intelligence right now.

The Hadoop system has become more and more complicated from when it was first initialised but the core components are still job tracker, task tracker and worker for the

MapReduce part. Different roles are created on DxChain Network to solve the problem of coordination.

7.2 Map-reduce in DxChain Network

MapReduce is a centralised design system, in which the job tracker manages cluster resources and job scheduling. The task tracker is in each agent and manages tasks in the nodes as well as communicates with the job tracker. The process of how MapReduce works is illustrated in Figure 7.1.

DxChain Network is a decentralised system which triggers the difficulty of keeping real-time communications between two nodes in a P2P network. Therefore, the CSC saves task distributions. Any computation miner who is interested in working on the task could potentially take the task.

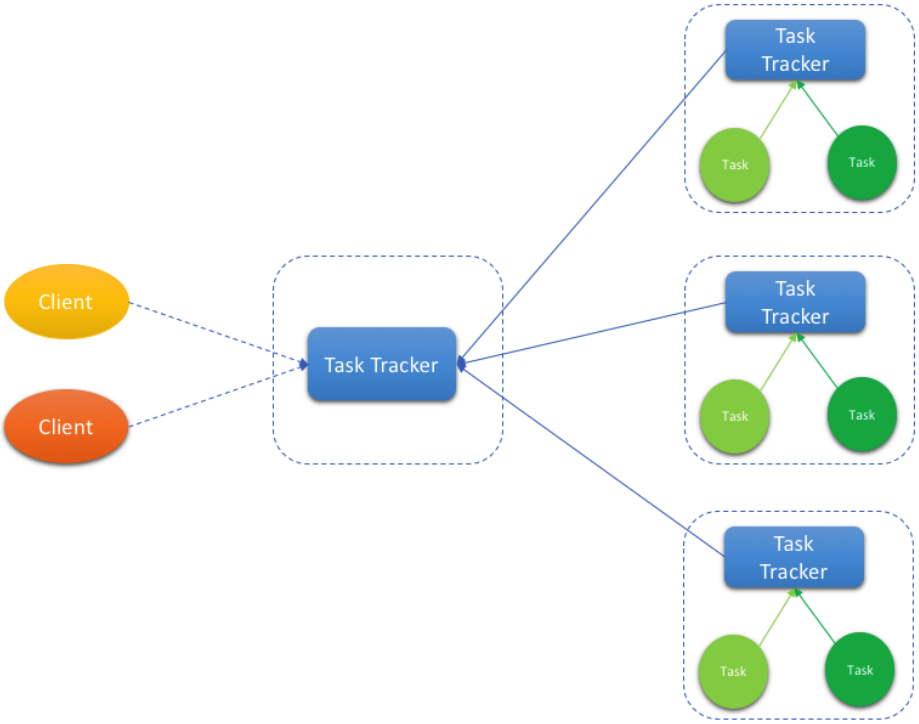

Figure 7.1: Illustration of general map-reduce process.

In the Hadoop system, besides distributing tasks to the task tracker, the job tracker communicates with the task tracker in order to know the activeness of the node through its activity status. If some of the nodes running active tasks are dead, the job tracker must re-assign tasks to new nodes. However, in DxChain Network, there is no need to check the statuses of the task nodes. More copies of redundant computations running in different nodes, as well as whether one or few nodes are off-line or dead, will not have an impact on the final result.

When a node finishes a computation, it sends the result to the CSC for verification through the verification game or PDC. The CSC saves job distribution information and results but these information are time-bounded, which allows the chain size to be shortened as it keeps only non-expired data.

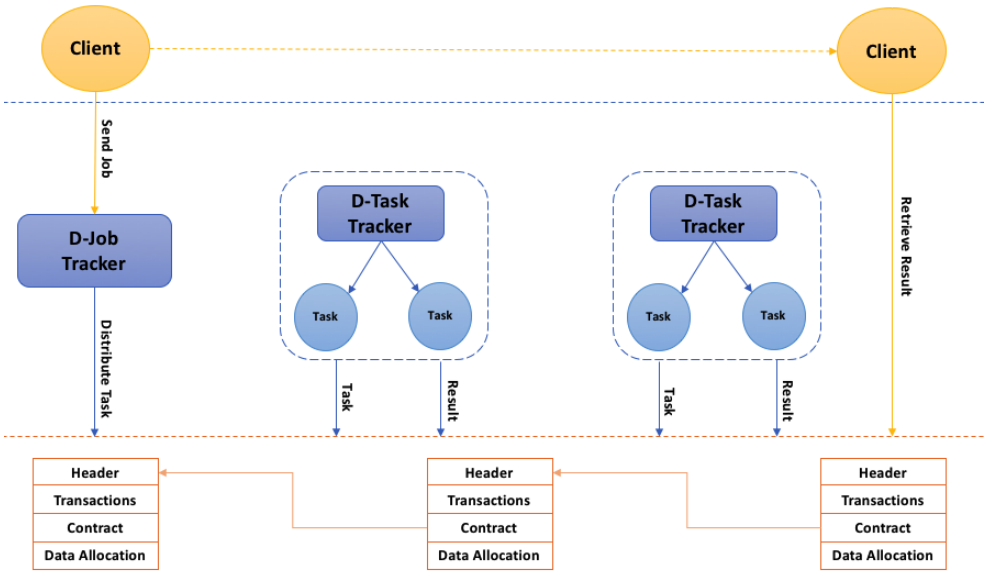

Figure 7.2: Illustration of map-reduce process in DxChain Network.

DxChain Network has two designated roles: D-Job Tracker and D-Task Tracker, to perform two different tasks. Figure 7.2 shows the MapReduce process in DxChain Network. A deposit which is calculated dynamically is needed to take up a role. The miner is intended to receive the incentive if the miner is honest in executing the task that it promises; otherwise the miner will lose the deposit. The incentives received through the computations will go to the main chain through the smart contracts.

7.3 Tool set in DxChain Network

Hadoop has a comprehensive ecosystem, including a huge collection of tools which could help easily run tasks such as Pig[5], Hive[3], HBase[2], Spark[6] and many more. HBase is a NoSQL database which needs a quick response for the query. DxChain Network is not compatible with HBase since the latency for the computation and consensus on the former is long, which does not fit the requirement. Spark provides in-memory computation which requires high memory machines so it is not an appropriate application in the DxChain Network. Pig and Hive do not have requirements of time-sensitivity and high memory machines, so they are good applications for DxChain Network.

Pig provides a set of high-level language for expressing data analysis. Pig supports database operations and analysis for non-structured data. If the data is stored as a plain text, Pig is a perfect tool to parse and analyse it on the fly.

Hive facilitates reading, writing and managing large datasets residing in distributed storage using SQL. DxChain Network supports database schema and SQL is used to do business intelligence related operations.

Mahout[4] is a framework to run machine learning algorithm including Canopy Clustering and Principal Component Analysis, etc. The DxChain team intends to select some of the machine learning algorithms to fit the platform of the DxChain Network.

The DxChain team intends to develop DPig, DHive and DMahout at this moment to facilitate the computation running on DxChain Network. More projects will potentially be coming.

Chapter 8

Privacy with Data Analytics

8.1 Introduction

Data analytics provides one with a large amount of prediction ability as one can notice trends and infer results. This is a central principle of the scientific method at a large scale when one can infer correlations. Even though correlation does not necessarily imply causation, strong correlation can often predict results. However, data analytics often requires inputs that contain personal data or data from individuals in a sample set. Predictions using inputs that come from individuals can be extraordinarily useful, however if done incorrectly, the private information of individuals can be exposed even from semi-anonymous gathering; with fully anonymous analytics, the quality of the data may be questioned and multiple datasets cannot be used together without using an identity of each input to compare that data. In addition to identities, if there is a fixed value between datasets that can compare each data point it may become a method to directly identify an individual.

When data is shared between datasets for analytics, the amount of information available is greater, hence allowing for some process of elimination if it is possible to query multiple datasets. To combat this, the DxChain team uses obfuscation techniques so that it is not possible to gather private information associated with individual users if rogue queries are made.

Even with perfect identity obfuscation, with enough values associated with a data-field that contains an identity, there is a possibility of back-tracing with the information and using some process of elimination, especially when knowing when/if a user has signed up. However, if the input of a user is not sufficient to change the results of a query into the

database, the responder to the survey has an expectation of privacy. This privacy can be quantified with differential privacy and can be used to strongly ensure that it would be nearly impossible to discern information about individual identities.

8.2 Encrypting Sensitive Information without private key disclosure

At times, sensitive information is stored in an Internet-accessible way as it may be required for identity verification (such as social security number when used for activities regulated by governing bodies). A major problem with trusting data encryption for "data at rest" is that encryption keys must exist on that server (or device). With asymmetric encryption, the ciphertext and public key may be stored, however if asymmetric encryption is used, it becomes possible to use a known plaintext attack. This is especially true with values with a small sample set such as social security numbers. With a public key and every possible social security number, it becomes easy to make a table of ciphertexts that correspond with various plaintexts and crack the values.

Therefore, if one is using asymmetric encryption, it is recommended to at least use padding, however in many cases that is not sufficient and the recommended option is to use a cipher like a Pallier cipher. During encryption, the ciphertext is multiplied by r^n where r is a random integer and therefore, there are multiple possible ciphertexts for every plaintext. The random number is removed when decrypting using Carmichael′s theorem. One other advantage of this method is that it is a homomorphic encryption method. When multiplying a ciphertext by a number x the corresponding plaintext increases by x

$$f(x) * A = f(x + a). \tag{8.1}$$

where $f(x)$ is a Pallier encryption. These techniques can be used for differential privacy to allow analytics results to change slowly as users add data and to prevent any major changes in query results.

8.3 Anonymous sharing of data

When sharing data between two databases, with differential privacy concerns aside, there may be a need to total up the number of participants with matching traits. For example, how many participants who had one disease in survey A also have the other disease. In

many cases, there is a concern when sending any information tied to an identity. There are many methods for matching survey participants individually while anonymously comparing data and still maintaining unique identities between participants. One example is with elliptic curve point multiplication. Consider $aP = Q$ where P and Q are points on an elliptic curve. Solving for a requires a solution to the elliptic curve discrete log problem. If each user is matched to a point on an elliptic curve and the keys are integers chosen in such a way that proper randomness is attained, it becomes possible to have one side encrypt an identity, send the encrypted identity to the other party and then have identities that are made anonymous

$$abP = baP = Q. \tag{8.2}$$

One potential attack may be that if the results are ordered, when information is shared, it may be possible to find users based on the order they are sent and received, therefore the list should be randomised when each side encrypts the users. This way 2 parties sharing results from 2 surveys with overlapping identities can share information. As more information is available to each survey because of the ability to query another survey, differential privacy methods are recommended in such a case.

8.4 Differential Privacy

When an individual enters data into a database, the results of any query can change. Changes to query results can be traced back to new entries in the database and make it easier to determine those individual identities with certain answers to a survey (or more broadly have a certain attribute stored in the database associated with the said user). Each person entering information stored in a database must have a method of denying that they were responsible for the changes. This becomes difficult when it is also important to protect the integrity of the data for proper analytics.

Differential privacy methods mitigate the probability of one user skewing query results and allowing information to be traced back to that user. Differential privacy depends on the user having deniability that he/ she was the one who changed the results.

8.5 Usage in DxChain Network.

The above four sections explain the importance of privacy and having some empirical way to keep privacy. However, there are no ways to solve the problem of privacy preserving

computation for large-scale general-purpose computation, such as MapReduce. DxChain Network will hence take into account privacy from the five aspects listed below.

Data Model Enabled Privacy Preserving. The DxChain Network supports data models for structured datasets, so clients should encrypt columns containing sensitive data such as SSN before submitting their data to the network. The searchable columns should not be encrypted. The reason is quite straightforward encrypted data must be decrypted before any computation is executed on it. For binary datasets, such as MP3, clients should encrypt the entire file.

Differential Privacy. If the users want to provide data only for statistical analysis, such as calculating mean and standard deviation, DxChain Network has a tool to facilitate the users in running differential privacy before submitting files to the network. After differential privacy is run, queries will not make sense since the entries have been modified via the addition of noise.

Miner Storage Encryption. The data piece is encrypted through using a storage miner public key in each local machine. Doing this protects against intrusion from network hackers as they do not know the private key of the miner.

Small Piece File. A large file will be split into many small pieces using different strategies, as only gaining access to a small fragment of a big file would not disclose much information.

Encryption During File Transfer. Before a file is copied to a storage miner, it is encrypted using the public key of the miner. This could protect its security through the traffic on the network.

The rule of thumb in DxChain Network is that if clients desire for their data to be stored privately, they must encrypt their data before submitting them to the network. The DxChain team considered privacy preserving computation using multi-party computation (MPC)[36], secret sharing[32], homomorphic encryption[30] and SGX[7]; they are all deemed to be not practical for the DxChain Network. As DxChain Network provides a big data and machine learning network, privacy preservation will be the responsibility of the users.

Chapter 9

DxChain Network-powered Ecosystems

The blockchain of the DxChain Network supports a variety of machine learning and data mining algorithms, data storage, music/ video streaming and many other applications. Developers from different industries could potentially leverage on the DxChain Network in different ways. This section describes a few usage scenarios in the DxChain Network-powered ecosystem.

9.1 Automated Data Sample Collection and AI Model Training

Artificial Intelligence (AI) has dramatically changed each sector of the global economy including advertising, finance, healthcare, transport, consumer, automation, energy, logistics and aerospace. By 2025, AI software and service revenue is projected to reach $59.8 billion worldwide[34]

As an AI technology driven security company, the quality of the malware and ransomware detection engine of the company Trustlook[8] heavily depends on the sample data collected from its customers and partners. Although Trustlook already has a solution to purchase or exchange its sample data, there are still several drawbacks:

1. High quality samples are difficult to obtain. Currently, the way to acquire samples is to purchase or exchange from large security vendors like Mcafee, Symantec and

Google. A startup company like Trustlook needs to pay a premium subscription fee to obtain samples from these companies. Many high-quality samples owned by these vendors are never available to the market to favour their own detection engines.

2. The cost to maintain a large machine learning cluster and data storage centre is high. The total size of mobile samples that Trustlook has collected has over several PB (Petabytes). When Trustlook extends its business from mobile to a more general network security field, the storage can easily grow to over ten times as large as what it is now. To maintain such a large data centre, budget and management will be big hurdles.

DxChain Network, a decentralised big data and machine learning blockchain, could potentially let the AI vendors benefit in many ways. The ecosystem developer could potentially leverage on the DxChain Network to build its own machine learning Dapp. Some of the advantages that DxChain Network could potentially provide are listed below:

1. Customised data models enable fully automated data collection through its malware detection Dapp defining the data collection format. The owner of each mobile device will define smart contract parameters, such as what kind of samples will be sent to the blockchain and at what price users would like to trade. Trustworthy sample data automatically collected through each device is sent to the blockchain of the DxChain Network.

2. Dramatically reducing data storage and network traffic cost, DxChain Network keeps its data blocks in the disk of each blockchain miner and each miner shares its own Internet bandwidth.

3. Decentralised machine learning algorithms are designed to be trained on a daily basis for the Trustlook AI model in the chain. This is just like shared economies, which can be used for almost free to share the unused CPU time of miners.

4. Immutable encrypted data on the blockchain protects the data of users, which is stored and presented in a very secure way without any potential worries about data breach or private data leak.

5. Each device user leverages on the smart contract of DxChain Network to make transactions and trade data in a trusted way.

Figure 9.1: Illustration of DxChain Network for Smart City.

9.2 Smart City

Smart City is a set of intelligent solutions to provide convenience for its residents at an infrastructural level and the market is likely to be worth a cumulative $1.565 trillion by 2020[33], which includes smart energy, smart building, smart mobility, smart technology, smart infrastructure, smart governance as well as smart education, smart security and smart citizens. Although much of the daily lives of people depend on smart cities, there are several drawbacks with the existing centralised networks:

1. Smart cities generate much more data than what the existing infrastructure can store and analyse. Based on a report that one trillion sensors will be deployed by [10], and that those sensors will help people to measure temperature, traffic patterns, foot traffic, air quality and security level of the infrastructure, considering that one sensor generates 10K of data, data containing inputs from one trillion sensors is a huge dataset. Another example is that a surveillance smart camera device using MPEG-4 with 30fps and 1080p resolution will fill the capacity of a 3TB drive in just 24 hours.

2. Without data mining, the collected data is useless but to operate a large data centre to cache and mine the data is beyond the ability of normal enterprises.

The users of the DxChain Network can potentially run machine learning data analysis directly on the dataset which the sensors generate each day and generate actionable business intelligence reports to guide the daily decisions of their businesses. The ecosystem

developer could potentially leverage on it to build its own machine learning Dapp. Below is a list of some of the advantages DxChain Network could potentially provide:

1. The sensor data collection of smart cities is automated via enabling the related smart contract in the Dapp by the developer.
2. Smart contracts intelligently control the running operations of all kinds of smart city sensors based on the surrounding environment, such as air conditioning temperature dynamic control to save more electrical power.
3. A chain data buyer could leverage on device data to develop a machine learning model to diagnose the running status of the device and predict device failure, in order to notify its vendor or schedule repair services ahead of time.
4. A chain developer could develop a machine learning model to help electricity companies such as PGE to optimise its scheduling model through generating the electricity consumption report and predicting the peak time electricity usage.

9.3 Healthcare

The valuation of the global healthcare market is projected to increase from US$31.71 billion in 2016 to US$57.85 billion by 2023 [29]. As the healthcare industry continues to grow bigger and bigger, the future of healthcare and Internet of Things are intertwined. The healthcare industry with the aid of smart devices provides on-the-diagnostics capabilities for people who cannot afford to go to hospital. Although the healthcare system has become better, more efficient and more effective, there are several drawbacks:

1. Healthcare premium is very high and many low-income people still cannot afford it.
2. It is difficult to integrate devices due to the phenomenon of healthcare IoT industry fragmentation.
3. There is no secure way that enables healthcare vendors to integrate and exchange data with other companies.

DxChain Network, a decentralised big data and machine learning network, could potentially enable the ecosystem developer to leverage on it to build its own big data and machine learning platform.

1. With DxChain Network, the cost for storing data and managing traffic will be reduced a lot compared to the traditional way, which will potentially help to lower the healthcare premium.

2. The data collected through fitness trackers, mobile apps, smart watches and other devices linked to the network are encrypted and stored on the blockchain in a tractable and secure way.

3. Data exchange and sharing is enabled by each vendor opening their APIs to others through using the data model of DxChain Network to standardise data.

4. Machine learning over the data of each individual device could potentially enable healthcare vendors to build their own AI technologies to monitor the health of patients and to send in critical vital signals to the community for the prevention of catastrophic events.

5. Finally, the ecosystem developer of the DxChain Network could potentially build on-device AI technology which can enable devices to talk to one another and materialise personalised treatment suggestions based on the various data points shared across smart devices.

Chapter 10

Risks

You acknowledge and agree that there are numerous risks associated with purchasing DX, holding DX, and using DX for participation in the DxChain Network. In the worst scenario, this could lead to the loss of all or part of the DX which had been purchased.

10.1 Uncertain Regulations and Enforcement Actions

The regulatory status of DX and distributed ledger technology is unclear or unsettled in many jurisdictions. The regulation of virtual currencies has become a primary target of regulation in all major countries in the world. It is impossible to predict how, when or whether regulatory agencies may apply existing regulations or create new regulations with respect to such technology and its applications, including DX and/or the DxChain Network. Regulatory actions could negatively impact DX and/or the DxChain Network in various ways. The Foundation (or its affiliates) may cease operations in a jurisdiction in the event that regulatory actions, or changes to law or regulation, make it illegal to operate in such jurisdiction, or commercially undesirable to obtain the necessary regulatory approval(s) to operate in such jurisdiction. After consulting with a wide range of legal advisors and continuous analysis of the development and legal structure of virtual currencies, the Foundation will apply a cautious approach towards the sale of DX. Therefore, for the token sale, the Foundation may constantly adjust the sale strategy in order to avoid relevant legal risks as much as possible. For the token sale the Foundation is working with Tzedek Law LLC, a boutique corporate law firm in Singapore with a good reputation in the blockchain space.

10.2 Inadequate disclosure of information

As at the date hereof, the DxChain Network is still under development and its design concepts, consensus mechanisms, algorithms, codes, and other technical details and parameters may be constantly and frequently updated and changed. Although this white paper contains the most current information relating to the DxChain Network, it is not absolutely complete and may still be adjusted and updated by the DxChain team from time to time. The DxChain team has no ability and obligation to keep holders of DX informed of every detail (including development progress and expected milestones) regarding the project to develop the DxChain Network, hence insufficient information disclosure is inevitable and reasonable.

10.3 Competitors

Various types of decentralised applications are emerging at a rapid rate, and the industry is increasingly competitive. It is possible that alternative networks could be established that utilise the same or similar code and protocol underlying DX and/or the DxChain Network and attempt to re-create similar facilities. The DxChain Network may be required to compete with these alternative networks, which could negatively impact DX and/or the DxChain Network.

10.4 Loss of Talent

The development of the DxChain Network depends on the continued co-operation of the existing technical team and expert consultants, who are highly knowledgeable and experienced in their respective sectors. The loss of any member may adversely affect the DxChain Network or its future development. Further, stability and cohesion within the team is critical to the overall development of the DxChain Network. There is the possibility that conflict within the team and/or departure of core personnel may occur, resulting in negative influence on the project in the future.

10.5 Failure to develop

There is the risk that the development of the DxChain Network will not be executed or implemented as planned, for a variety of reasons, including without limitation the event of a decline in the prices of any digital asset, virtual currency or DX, unforeseen technical difficulties, and shortage of development funds for activities.

10.6 Security weaknesses

Hackers or other malicious groups or organisations may attempt to interfere with DX and/or the DxChain Network in a variety of ways, including, but not limited to, malware attacks, denial of service attacks, consensus-based attacks, Sybil attacks, smurfing and spoofing. Furthermore, there is a risk that a third party or a member of the Foundation or its affiliates may intentionally or unintentionally introduce weaknesses into the core infrastructure of DX and/or the DxChain Network, which could negatively affect DX and/or the DxChain Network.

Further, the future of cryptography and security innovations are highly unpredictable and advances in cryptography, or technical advances (including without limitation development of quantum computing), could present unknown risks to DX and/or the DxChain Network by rendering ineffective the cryptographic consensus mechanism that underpins that blockchain protocol.

10.7 Other risks

In addition, the potential risks briefly mentioned above are not exhaustive and there are other risks (as more particularly set out in the Terms and Conditions) associated with your purchase, holding and use of DX, including those that the Foundation cannot anticipate. Such risks may further materialise as unanticipated variations or combinations of the aforementioned risks. You should conduct full due diligence on the Foundation, its affiliates and the DxChain team, as well as understand the overall framework, mission and vision for the DxChain Network prior to purchasing DX.

Chapter 11

Future Research Work

Since research in the areas of blockchain technology and consensus protocol is very active, the DxChain team aims to explore more Proof of Useful Work, Proof of Storage and Proof of Computing schemes. As DxChain Network uses several dynamic adjustments, such as incentives, mining difficulties and role elections, the researchers of the DxChain team intends to conduct more research on finding more stable parameters and mechanisms for those adjustments.

Chapter 12

Conclusion

In this White Paper, the DxChain Network is introduced— a scalable, private and extensible blockchain dedicated to big data and machine learning, with its architecture and core technologies including: 1) a new decentralised computing framework to introduce Provable Data Computing and verification game; 2) a new chains-on-chain design to orchestrate a master chain and two side chains— Data Side Chain (DSC) and Computing Side Chain (CSC); 3) Hadoop incorporation to DxChain Network to facilitate big data and machine learning and 4) flexible and powerful DxChain Network-based system architecture, dedicatedly designed to support most business data exchange and data analytics requirements.

Chapter 13

Acknowledgements

This work is the cumulative effort of multiple individuals within the DxChain team; it would not have been possible without the help, comments and reviews of the collaborators and advisors of the DxChain team. The DxChain team would like to express its gratitude to its contributors, advisers and to the many people in big data, distributed computing, IoT and other cryptocurrency communities for their early reviews and constructive suggestions.

References

[1] *Apache Hadoop*. http://hadoop.apache.org/.

[2] *Apache HBase*. https://hbase.apache.org/.

[3] *Apache Hive*. https://hive.apache.org/.

[4] *Apache Mahout*. https://mahout.apache.org/.

[5] *Apache Pig*. https://pig.apache.org/.

[6] *Apache Spark*. https://spark.apache.org/.

[7] *SGX*. https://software.intel.com/en-us/sgx.

[8] *Trustlook*. https://www.trustlook.com.

[9] *The Golem Project*. 2016. http://golemproject.net/doc/DraftGolemProjectWhitepaper.pdf.

[10] *Over 1 trillion sensors could be deployed by 2020*. 2016. https://www.electronicspecifier.com/around-the-industry/over-1-trillion-sensors-could-be-deployed-by-2020.

[11] Juan Benet. *InterPlanetary File System (IPFS)*. 2014. https://github.com/ipfs/papers.

[12] Juan Benet. *Filecoin: A Cryptocurrency Operated File Storage Network*. 2017. https://filecoin.io/filecoin-jul-2014.pdf.

[13] Vitalik Buterin. *Ethereum*. 2013. https://www.ethereum.org/.

[14] Melissa Chase and Anna Lysyanskaya. *On Signatures of Knowledge.* 2006. https://iacr.org/archive/crypto2006/41170076/41170076.pdf.

[15] Dominik Schiener David Sønstebø, Sergey Ivancheglo and Dr. Serguei Popov. *IOTA.* 2015. https://blog.iota.org/.

[16] Jeffrey Dean and Sanjay Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters.* 2004. https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

[17] Jeffrey Dean Fay Chang and et al. *Bigtable: A Distributed Storage System for Structured Data.* 2006. https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf.

[18] Randal Burns Giuseppe Ateniese and et al. *Provable Data Possession at Untrusted Stores.* 2007. https://people.eecs.berkeley.edu/~dawnsong/papers/p598-ateniese.

[19] Ari Jakobsson, Markus; Juels. *Proofs of Work and Bread Pudding Protocols.* Communications and Multimedia Security. Kluwer Academic Publishers: 258–272., 1999. .

[20] Christian Reitwießner Jason Teutsch. *A scalable verification solution for blockchains.* 2017. https://people.cs.uchicago.edu/~teutsch/papers/truebit.pdf.

[21] Vitalik Buterin Joseph Poon. *Plasma: Scalable Autonomous Smart Contracts.* 2006. https://plasma.io/plasma.pdf.

[22] Sanjay Radia Robert Chansler Konstantin Shvachko, Hairong Kuang. *The Hadoop Distributed File System.* 2014. http://storageconference.us/2010/Papers/MSST/Shvachko.pdf.

[23] IoTex Lab. *IoTex: A Decentralized Network for Internet of Things (IoT).* 2017. https://iotex.io/.

[24] Protocol Labs. *Proof of Replication Technical Report (WIP).* 2017. https://filecoin.io/proof-of-replication.pdf.

[25] Manuel Sabin Prashant Nalini Vasudevan Marshall Ball, Alon Rosen. *Proofs of Useful Work.* 2016. https://eprint.iacr.org/2017/203.pdf.

[26] Camille Marini Mathieu Galtier. *Traceable Machine Learning on Hidden data*. 2017. https://arxiv.org/pdf/1704.05017.pdf.

[27] mthcl. *The math of Nxt forging*. 2014. https://www.docdroid.net/e29h/forging0-5-1.pdf.

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

[29] Transparency Market Research. *Smart Healthcare Products Market*. 2016. https://www.transparencymarketresearch.com/smart-healthcare-products-market.html.

[30] Adam L.; Yung Moti. Sander, Tomas; Young. *Non-Interactive CryptoComputing For NC1*. FOCS1991. IEEE. .

[31] Howard Gobioff Sanjay Ghemawat and Shun-Tak Leung. *Google File System (GFS or GoogleFS)*. 2003. http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf.

[32] Adi Shamir. *How to share a secret*. Communications of the ACM. 22 (11): 612–613., 1979. .

[33] Frost Sullivan. *Strategic Opportunity Analysis of the Global Smart City Market*. http://www.egr.msu.edu/~aesc310-web/resources/SmartCities/Smart%20City%20Market%20Report%202.pdf.

[34] Tractica. *Artificial Intelligence Software Revenue*. https://www.tractica.com/newsroom/press-releases/artificial-intelligence-software-revenue-to-reach-59-8-billion-worldwide-by-2025/.

[35] Pavel Vasin. *BlackCoin's Proof-of-Stake Protocol v2*. 2014. http://blackcoin.co/blackcoin-pos-protocol-v2-whitepaper.pdf.

[36] Andrew C. Yao. *Protocols for Secure Computations (extended abstract)*. 1982. http://research.cs.wisc.edu/areas/sec/yao1982-ocr.pdf.