A Blockchain and Directed Acyclic Graph-integrated Secure Data Flow System

Jack Ge[1] **Ansel Gao**[2] **Arthur Yu**[2]

Jerry Ning[2] **Giorgi Javrishvili**[2] **Simon Xie**[2]

[1]Zhejiang University, [2]CyberVein Research - Hangzhou

Abstract

In today's era, data is being created in an astonishing pace, but data in many areas is limited in quantity and of poor quality. Some people have made estimates that if medical data was marked by a third-party company, it will take 10,000 people 10 years to collect valid information. In most industries, due to factors such as competition, security issues, and approval processes It is almost impossible to integrate decentralized data, or the cost is huge. Therefore, there is a barrier that is difficult to break through the flow of data, the so-called "data island" problem. Even if the industry intentionally exchanges data, it may also encounter policy accountability, because the emphasis on data privacy and security has become a worldwide trend. How to protect user privacy and prevent sensitive information leakage in the process of big data development and application has become a new challenge. In response to the above problems, CyberVein adheres to an open, shared, and co-constructed attitude, taking blockchain technology as the core to integrate data in various aspects and angles, and low-cost decentralized construction and participation through the participation of more nodes maintain. At the same time, CyberVein also combined with the database algorithm framework of personal terminal devices, and proposed a systematic and general solution based on "federal learning", which can solve the problem of joint modeling of individuals (to C) and companies (to B). It can realize that the own data of the C-side and B-side does not go out of the local, but through the parameter exchange method under the encryption mechanism, a virtual shared model is established without violating the data privacy regulations. Because the data itself does not move, it does not involve privacy leaks and data compliance issues, and the models built only serve their respective regions. Under such a mechanism, the identities and statuses of all parties involved are equal, and token rewards are obtained by contributing their own data to model, successfully achieving the goal of "common prosperity".

Design principles and concepts

Industry background

Data leakage, domestic and international attention to network security

Although the Internet has been popular for many years, but the public's information security awareness is still not strong enough, and online fraud, personal privacy disclosure, and theft of commercial data are only increasing. Numerous citizens provide data invisibly and free of charge, and the data is controlled by a very small number of people invisibly to avoid the imbalance and conflict of the international community caused by the contradiction of data resources. In June 2017, China began to implement the "Network Security Law of the People's Republic of China", stating that "network operators must not disclose, tamper with, or damage the personal information they collect; they must not disclose the information to others without the consent of the information provider." In May 2018, the European Union introduced the first policy for data privacy protection, the General Data Protection Regulation (GDPR), which clarified certain provisions on data privacy protection. In October 2019, the Standing Committee of the National People's Congress voted to pass the cipher law of the People's Republic of

China, marking that the password will become the core technology to protect China's cyberspace security and the core position of cyber security. This series of initiatives means that the collection of user data must be open and transparent. Data cannot be exchanged between enterprises and organizations without user authorization. At the same time, it also guarantees network and information security, and protects data privacy and data ownership under the framework of risk regulation and consumer protection.

Data explosion, the promotion of big data platform upgrade

On October 31, 2019, China's three major telecom operators announced the price of 5G on the same day, which marked the official entry of the commercial use of China's 5G network. With the high bandwidth, low latency and AI capabilities of 5G, new applications such as Internet of Vehicles, intelligent manufacturing, smart energy, wireless medical, wireless home entertainment, and drones have been created, creating new and rich data dimensions, and the business forms they carry are more complex and diverse. In view of the problems of data concentration, large data volume and sparse data value, in the business development, a single computing platform and a centralized data platform are difficult to focus on valuable data. Therefore, whether it can effectively deal with complex, diverse and massive data collection, processing and interaction has become a huge challenge, and a new, decentralized big data platform has also become a development trend.

Data resources, promotion of blockchain technology integration and innovation

With the “high-level adjustment of the blockchain as a core technology for independent innovation” and “accelerate the promotion of blockchain technology and industrial innovation development”, the development and application of blockchains have risen to national strategies. The 21st century is the century of data, and data is a strategic resource. When data becomes an asset, it means protection, prevention of copying and tampering. Changes in data attributes have placed new demands on various technologies! This is also an important driving force for the rise of blockchain technology. However, blockchain technology is not a panacea. As a new generation of information technology, blockchain must be integrated with other technologies to maximize the economic and social benefits. Blockchain’s integration innovation and convergence with artificial intelligence, cloud computing, Internet of Things and big data technology has become the general trend. Therefore, any high-quality blockchain project should have the characteristics of privacy protection, fair competition, data interaction and technology integration innovation.

Visions

Today, issues such as “small data”, “data silos”, “data breaches” and “unknown data ownership rights” have hindered the digital economy and the commercialization of technology. CyberVein will follow the principles of using technology for the good and the general public to solve the existing data problems.

Mining data value

The discovery of data value is different from ordinary commodities. Only by deeply understanding the characteristics of data in technical design and mining of data can we continuously improve the ecology of data value discovery. For example, passively generated data (such as cookies) has no value in itself, and only has value in exchange when it has usage value; the exchange value of data needs to be defined by the carrier and medium, only then can the value of data be quantified and data can be transferred. The ownership of data is the ability to define in detail of how the data owners distribute, share, and deliver data to other parties, it has rights to define data owners and third-party users to create, edit, modify, share, and restrict access to the data.

Realizing the commercial application of data

In the information age, the in-depth analysis and application of big data made everyone realize the importance of data, and how to legalize and capitalize the massive amount of data has become the most relevant issue for the public and various industries. After an in-depth study of blockchain technology and data value, CyberVein has created a global blockchain service infrastructure that spans across the chain, across geographies, and across organizations. Anyone using any smart device on any chain can use data modeling, by contributing and sharing data, by realizing data value and more efficient management, achieving higher value conversion of data at a lower cost.

Analysis and innovation of technology

The value of blockchain technology has been gradually tapped by society, but in the commercial, government and civilian fields, most blockchain technologies still have many hidden dangers and flaws, such as limited data storage capacity, long transaction time and low transaction efficiency, increase in the difficulty of node expansion accounting, consensus mechanism leading to irrational resource allocation, limited block capacity and limitations in various application areas. Therefore, CyberVein focuses on improving the shape and function of existing blockchains, and integrates technologies such as artificial intelligence (AI), 5G, Internet of Things (IoT) and big data. Its effective cross-chain technology can also more effectively and securely build a platform that can define and carry the value of data, providing important support for the development and operation of upper-layer applications of the blockchain.

Design Principles

For current economic and social issues, CyberVein achieves its goals through the following solutions:

Decentralization, emphasizing on personal data rights

CyberVein strictly follows the logic of decentralization, providing only the federated learning's operating environment and open source framework, which can define in detail the ability of data owners to distribute, share and deliver the data to other parties. Your data ownership is in your hands and the data is only kept locally. Data can be modeled automatically only if you authorize it.

Data security of the public

CyberVein meets the needs of user privacy protection and data security. The data cannot be removed locally, and the parameters of the model are not only encrypted when processed by the third party but ensures that any characteristics of the original user cannot be reversely analyzed, achieving the fair encrypted exchange of information and model parameters.

Efficient interaction, paying attention to each data quality

CyberVein guarantees that the data interaction is effective in ensuring that the quality of the model is not lossless, and there is no negative migration, ensuring that the federal model is better than the split independent model. And the use of DAG architecture as the underlying technology, can support tens of millions of users, in theory, there is no throughput cap.

Open source sharing, paying attention to all data rules

It is a decentralized federated learning operating environment and open source framework. Anyone can write a database and publish smart contracts and DApp, so they can develop the rules for data ownership, data transaction formats, and data value conversion capabilities in their own projects.

Product architecture and technical solutions

Protocol layer

CyberVein's basic protocol is divided into the DAG layer and the blockchain layer, which correspond to the data storage layer and the data management and value transaction layer, respectively. The two communicate and connect through super-nodes. Such a two-layer design not only guarantees high scalability and transaction throughput of mass data, but also guarantees atomicity and uniformity of transaction status (Transaction), and also ensures that crypto assets are protected from malicious attacks and double spending during the exchange process.

DAG layer (directed acyclic graph)

DAG is a directed acyclic graph. In the network based on block and chain structure, due to the limitations of its design, there are problems such as low throughput, slow transaction confirmation time, and node data expansion. Compared with the blockchain, DAG mainly promotes synchronous accounting to asynchronous accounting, which can greatly improve scalability. A distributed database using DAG technology can initially achieve TPS of 100,000+, and theoretically the throughput has no upper limit.

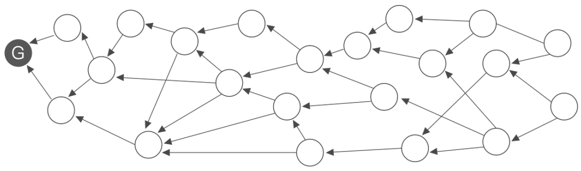


Figure 1: Schematic diagram of the DAG structure.

In DAG, there is no concept of a block. His constituent unit is a single transaction, and each unit records a single user's transaction. Compared with the chain structure, the time for packaging blocks is saved, and the method of verifying transactions Relies on the verification of the previous transaction by the latter transaction.

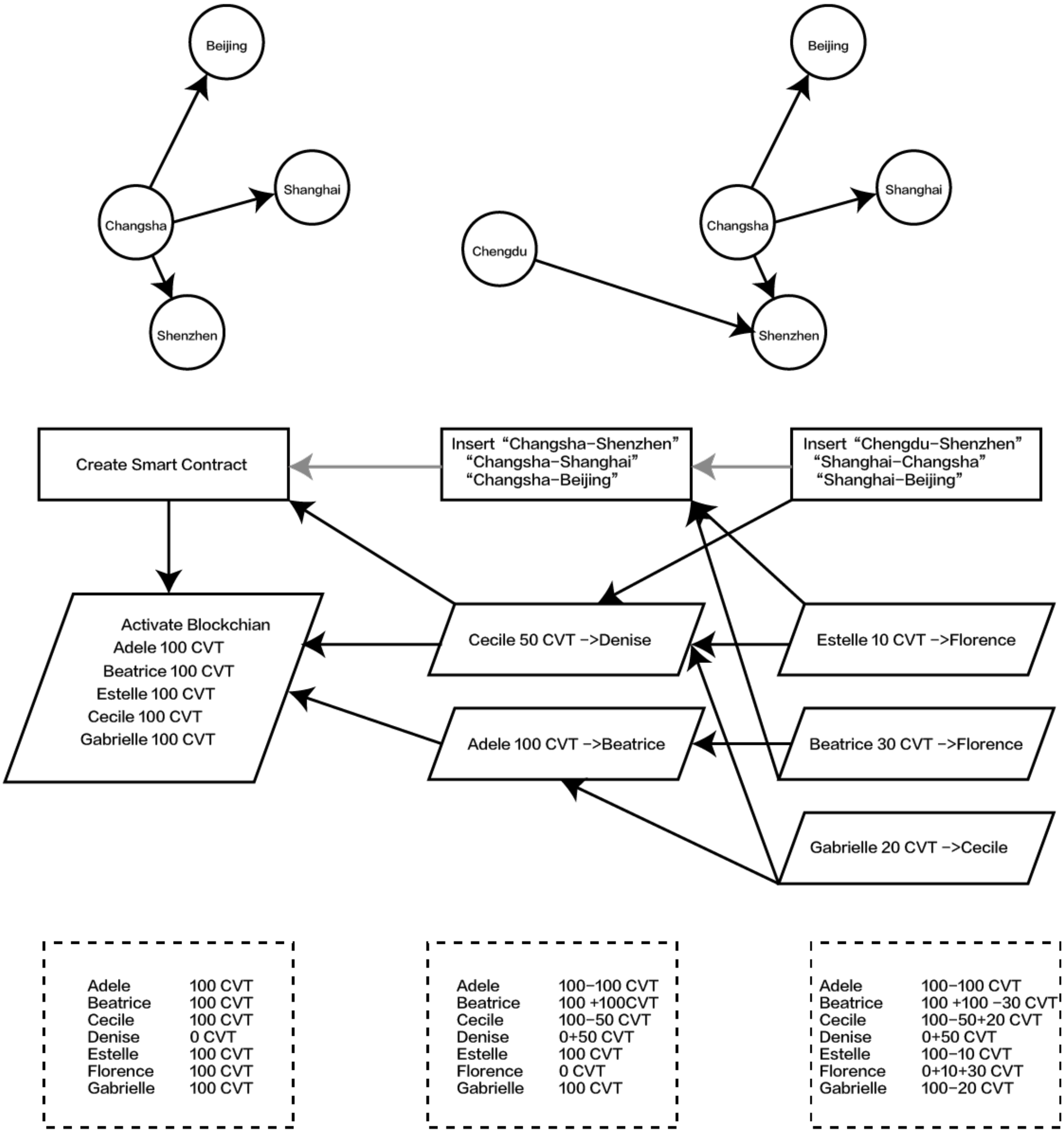
In CyberVein, data is written into a unit as a transaction, and the verification method is that different nodes undertake different tasks.

Blockchain Layer

DAG's asynchronous communication mechanism has obvious advantages in improving scalability, shortening confirmation time, and reducing payment fees, but its security and consistency issues are difficult to properly resolve. Therefore, CyberVein puts smart contracts and CVT on blockhain to help Ensure the security of transactions and the effectiveness of smart contract processing.

In CyberVein, the quantification of data, data flows and their access rights (CVT payment) is implemented by means of smart contracts (ie database (set) management). Databases (sets) can be accessed only through corresponding smart contract functions, which include paying a certain amount of CVT fees and network transaction processing fees.

CyberVein combines two methods of storing large amounts of structured data. The state of the database (set) is the sum of the data applied in



Adele	100 CVT
Beatrice	100 CVT
Cecile	100 CVT
Denise	0 CVT
Estelle	100 CVT
Florence	100 CVT
Gabrielle	100 CVT

Adele	100–100 CVT
Beatrice	100 +100CVT
Cecile	100–50 CVT
Denise	0+50 CVT
Estelle	100 CVT
Florence	0 CVT
Gabrielle	100 CVT

Adele	100–100 CVT
Beatrice	100 +100 –30 CVT
Cecile	100–50+20 CVT
Denise	0+50 CVT
Estelle	100–10 CVT
Florence	0+10+30 CVT
Gabrielle	100–20 CVT

Figure 2: Blockchain data management and DAG data storage authentication process and value transaction process.

order and all data modifications. These modifications are packaged into smart contracts that belong to the set of representative data (libraries). Smart contract transactions that include data set updates only affect nodes that maintain the local database of the data set and update the local database by executing modification instructions. This makes CyberVein's management of the database on the one hand prevent fraud and tampering, and the data is stored intact; on the other hand, the need to continuously add and modify data in the database can be effectively met.

The middle row of the figure above shows an example of the CyberVein blockchain. The diamond-shaped box contains the CVT transactions in CyberVein, and the bottom line shows their impact on the accounting status. The directed acyclic in the middle row also contains smart contract transactions (rectangular boxes) and data set modifications. The top line shows the impact of data management (ie, smart contracts in CyberVein) on the database (set).

The structure of the CyberVein network consists of two chains, the DAG network for data storage; and a smart contract integrated blockchain network for contributing verification consensus and data processing, while using the same consensus. Both networks are connected through ***validators*** (super nodes).

The blockchain network in CyberVein processes and records the digital asset transactions, which consists of a series of transaction blocks β_0, β_1, β_2,...which act as a public ledge of all transactions happening on the CyberVein network when a contributor initiates data exchange on the network. Every block $\beta_i = (\varphi_i, \sigma_i, \tau_i)$ is separated into three sub-blocks that contain the index i that represents the block position.

The transaction sub-block τ_i contains

- The current block sequence number i,
- A list of *transactions*.

The hash sub-block φ_i contains

- The current block sequence number i,
- The contributor's timestamps ζ_φ on φ_{i-1}, the $(i-1)$th HASH sub-block,

The SIGNATURE sub-block σ_i contains

- The current block sequence number i,
- The contributor's signature ζ_τ on τ_i, the (i)th TRANSACTION sub-block,
- The contributor's signature ζ_τ on σ_{i-1}, the $(i-1)$ SIGNATURE sub-block.

Consensus layer

One of the main challenges of blockchain technology, that is, open and distributed ledger systems, is network maintenance, because the network is usually open to everyone, including malicious actors. This requires the establishment of an economic model to punish malicious actors and reward those who support and maintain the network.

The conventional approach is to introduce some kind of economic barriers for participating in the consensus process, so that malicious actors cannot attack the consensus process for free. On the other hand, the rewards of participating in the consensus process need to exceed this economic barrier, otherwise the rewards and motivations for promoting good participation cannot exist. Generally, from the perspective of network participants, the barrier is to exchange scarce resources for platform access (the ability to write shared ledgers / records) and digital assets.

CyberVein uses a consensus algorithm, and scarce resources are potential for consensus participants to provide disk space. Providing disk space to store value and smart contract transactions is directly beneficial to the system's purpose of promoting distributed storage of structured data. In order to be eligible to charge transaction fees to users who place transactions on the network, consensus participants must demonstrate that they have stored transactions previously confirmed by the network.

Proof of Contribution (PoC) is the consensus mechanism used by CyberVein. Its core value is that the contribution of storage space and bandwidth on the nodes is provided to the entire network, generating the rewards (CyberVein Tokens) received by the nodes. Although this consensus mechanism is not a new concept, due to the excessive energy consumption of POW mining, PoC has become more and more popular worldwide in the development of blockchain technology. Based on the original PoSpace consensus, CyberVein aims to provide file storage and retrieval services by "using storage space mining", allowing real applications and more flexible development on the network while maintaining advanced energy-saving features.

Generally speaking, it is not the CPU time or other scarce resources that ultimately secure the network: it is the nodes in the network. Nodes will verify transactions and protect the information stored on the chain from malicious behavior by rejecting transactions that attempt to push consensus to an inconsistent state.

Nodes

There are basically two types of nodes: ordinary participants (basic nodes) with data upload, download, transfer, message synchronization, and storage, and implementation of DAG and blockchain cross-chain consensus calculation. At the same time, it is responsible for the processor (super node) that stores the verification records. This node can be an individual or a company.

In ordinary nodes,

Data uploader: After paying the upload fee (CVT), the data uploader can upload the data set and wait for the user to download. The download fee (excluding transaction fees) paid by the data downloader will be the main source of income for the data uploader.

Data downloader: The downloader needs to pay the download fee (CVT) after sending the download request. If the request is verified as valid by the supernode, the downloader will have access to the data set and download the data.

Contributors: Contributors provide cloud storage services for data and receive corresponding rewards through the contributed storage space and bandwidth.

The steps are shown below:

1. The user calls the api, stores the file in node 1, and signs it with his private key to initiate a payment transaction.

2. The user shard the file into $f_1 - f_n$ and generates N random numbers $r_i - r_n$, and calls the hash function$h(i) = hash(f_i, r_i)$, which $1 \leq i \leq n$ generates a taglist file (splitting the file location, the random number, the corresponding hash value, the private key signature and the public key), and stores it in the super node (Any one available).

3. After node 1 receives the storage and transaction request (at this time, the transaction is not yet confirmed), it looks for the current central super node to verify the storage certificate. Node 1, Node 1 returns the result based on its own file, user public key, random number in tag, and offset as proof. After the central node verifies proof, it returns success, and globally uniquely verifies the id and its own (super node) private key. signature.

4. Node 1 waits until the verification result is broadcast.

5. Other nodes verify the transaction (the central node's public key is used to verify the signature and the globally unique verification id), and write it into the DAG, and the transaction is determined when all subsequent transactions are covered.

This verification process is based on the Proof of Possession of Data (PDP) used to verify the availability of data and will be explained in detail in next section.

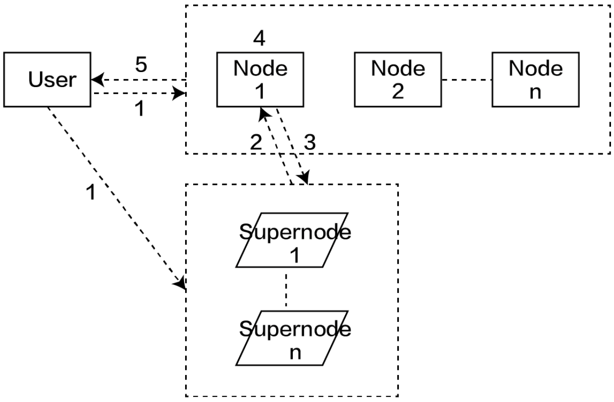


Figure 3: Normal node and super-node authentication process when users retrieve files.

Contribution verification

Instead of measuring hashing power in that of PoW, or the amount of stake held in the node like that of in PoSpace, CyberVein evaluates and decide the block to be added by the next eligible contributor on the network, which measures PoC of storage and bandwidth. The PoSpace from [14, 34] are specified a family of “hard-to-pebble” directed acyclic graphs of increasing size.

Algorithm 1: Storage and bandwidth commit

Common input: A hard-to-pebble graph G with n nodes and a function hash: $\{0,1\}^* \rightarrow \{0,1\}^L$.

1. P generates a unique nonce μ and then computes and stores $\left(\gamma, S_\gamma\right) := \text{Init}(\mu, n)$,

and sends the nonce μ and the commitment γ to V. S_γ contains the labels of all the nodes of G computed using Eq. (1) and γ is a Merkle-tree commitment to these n labels. The total size of S_γ is $N = 2 \cdot n \cdot L$ (graph + Merkle tree).

Algorithm 2: Prove commit

Initial state: V holds commitment γ and nonce μ;

P stores S_γ and μ. Both are given the challenges $c = \left(c_1, \ldots, c_{k_v}\right)$ to be used.

1. P computes openings $b := \left(b_1, b_2, \ldots\right)$ of all the labels of the nodes $\left\{c_i\right\}_{i \in [k_v]}$ and of *all their parents* and sends them to V. This is done using Ans where $\text{Ans}\left(\mu, S_\gamma, c\right)$ returns the Merkle inclusion proof of label l_c w.r.t. γ.
2. V verifies these openings using Vrfy , where $\text{Vrfy}(\mu, \gamma, c, a) = 1$ if a is a correct opening for c. It then checks for all $i = 1, \ldots, k_v$ if the label l_{c_i} is correctly computed as in Eq. (1).

Algorithm 3: Prove storage and bandwidth
Initial state: V holds commitment γ and nonce μ;

P stores S_γ and μ. Both are given the challenges $c = \left(c_1, \ldots, c_{k_p}\right)$ to be used.

1. P computes openings $\left\{a_i := \text{Ans}\left(\mu, S_\gamma, c_i\right)\right\}_{i \in \left[k_p\right]}$ and sends them to V.
2. V verifies these openings by executing $\text{Vrfy}\left(\mu, \gamma, c_i, a_i\right)$.

Each block at the specific time step records set m of valid *contribution* γ of storage and bandwidth that passed the threshold value j, where all sets of valid contributions $\pi_1 = \left(pk_1, \gamma_1, c_1, a_1\right), \ldots, \pi_m = \left(pk_m, \gamma_m, c_m, a_m\right)$ could win a corresponding proportion of reward based on ith contributor's fraction of the total contribution in the network. The lacking of either contribution of storage or bandwidth will lead to the reduction of $\Pr_{\text{hash}}$ to zero, this is proposed to reduce the possibility of useless contribution that serve neither storage nor upload/download purpose and ability. The process of hashing answer a is simulated through random oracle hash, as follows:

$$\Pr_{\text{hash}}\left[\forall j \neq i : \text{ Quality }\left(\pi_i\right) > \text{ Quality }\left(\pi_j\right)\right] = \frac{N_{\gamma_i}}{\sum_{j=1}^{m} N_{\gamma_j}} \frac{M_{\gamma_i}}{\sum_{j=1}^{m} M_{\gamma_j}} \tag{1}$$

where N_{γ_i} is the space committed to by γ_i, M_{γ_i} is the bandwidth committed to by γ_i.

Static PDP file verification

PDP-Proof of Data Possession is used to verify data availability, completeness and retrievability. In general, data storage outsourcing has many advantages, but comes with some risks. For example, a storage provider deletes a local copy of a data file after receiving a token reward. Based on this situation, the PDP scheme is sufficient to effectively check whether the remote cloud server has retained the data set and enable the storage provider to prove to the verifier that the save set is retrievable.

Basic framework

Because the entire data set is divided into many small data blocks and then stored, the entire package is not fully encrypted in the design of the entire system. Only a few bits of data are encrypted per data block, which significantly improves client efficiency by minimizing computation and storage overhead. [11]

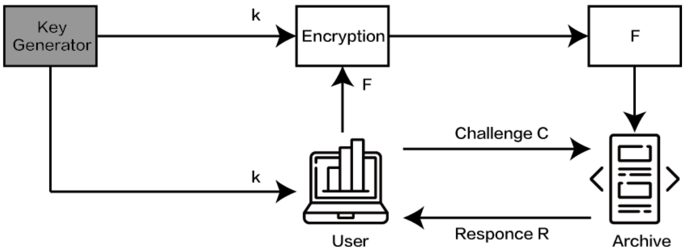


Figure 4: General description of the document retrieval challenge process.

It not only reduces the storage and computing overhead of the client and server, but also minimizes the scope of data integrity certification and the consumption of network bandwidth.

PDP for Large files

In CyberVein PDP scheme, the verifier/challenger stores only a single cryptographic key - irrespective of the size and number of the data whose retrievability it seeks to verify - as well as a small amount of dynamic state for each file. It's worth noting that PDP scheme allows that the prover (storage provider) access only a small portion of a large data file E; and in general, this small portion of E is essentially independent of the length of F. Briefly, this PDP protocol encrypts E and randomly embeds a set of randomly-valued check blocks called sentinels.[5] Here, the encryption algorithm makes the sentinel and other files indistinguishable. The verifier challenges the prover by specifying the location of a set of sentinels and asking the prover to return the relevant sentinel value. If the prover modifies or deletes most of F, then it is likely that it will also suppress some sentinels. Therefore, it is unlikely that the prover will respond correctly. In order to prevent a small part of the prover from corrupting, we also use error correction codes.

Smart Contract Layer

Unlike other smart contracts on the blockchain, Cybervein's contract layer consists of a database system maintained by multiple nodes.

The contract layer database system consists of the following important components:

User API interface. The contract system provides the user with an interface to operate the database in the form of a client.

The core service program Database service is used to provide core services such as interaction with clients, database data synchronization, local database access, and docking with Cybervein DAG. The core service programs of multiple nodes form a real-time data synchronization network and ensure Byzantine fault tolerance between nodes.

The Cybervein peer node is used to complete blockchain-related operations and provide an interface to access smart contracts on the blockchain.

Local database instance, the data of the database instance on each node is a complete backup of the data stored on the entire blockchain, and guarantees real-time updates.

For database storage, it is mainly divided into two parts, one is the basic data file, and the other is the incremental log file. The base data file is the full amount of data of a database at a time stored in binary form, which is called base db contract in this system. The other part of the incremental log file refers to the operation log of all operations performed on the database during the operation of the database system based on the base db contract. Then, the storage of the base db contract and the ordered operation log can completely store the entire amount of data in the entire database system at any time.

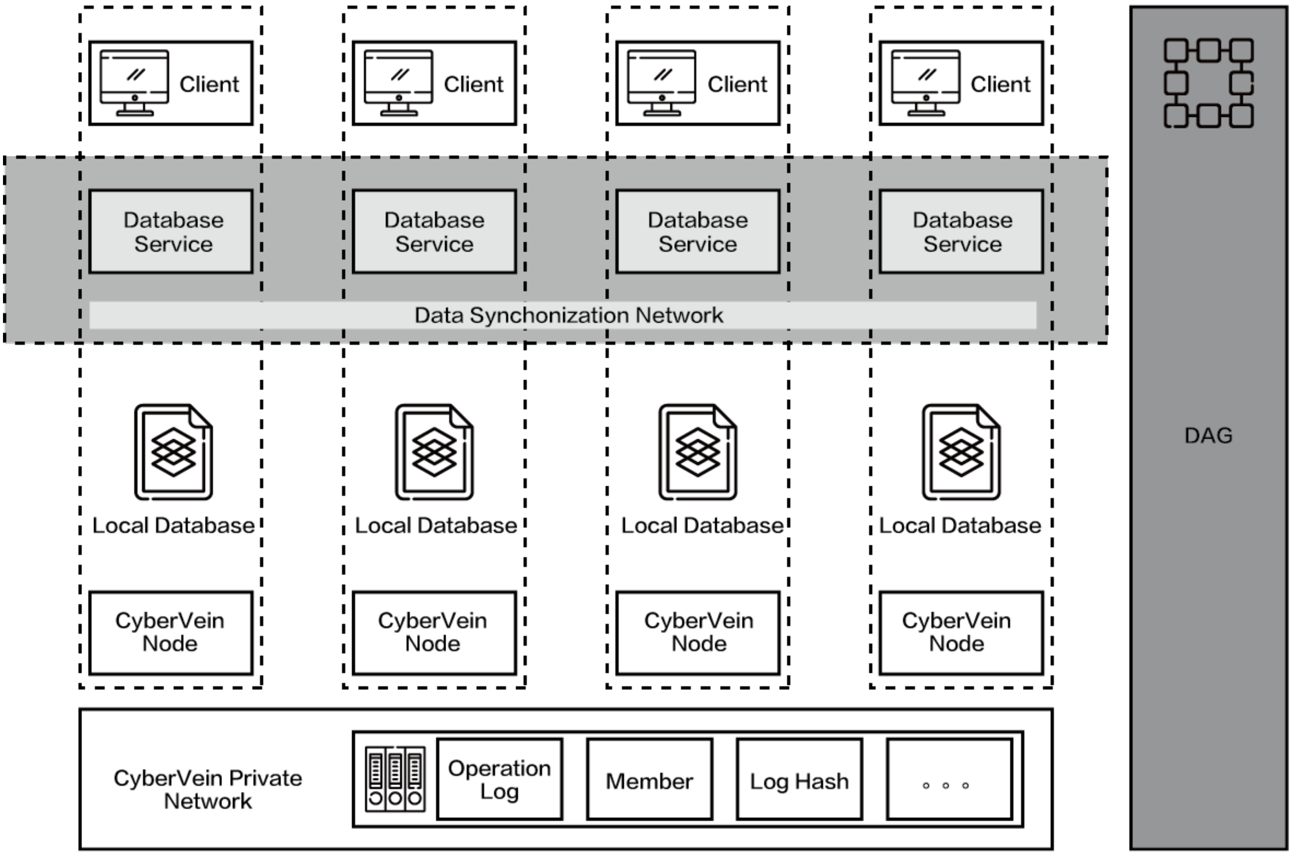


Figure 5: CyberVein blockchain smart contract description.

Smart Contract Design

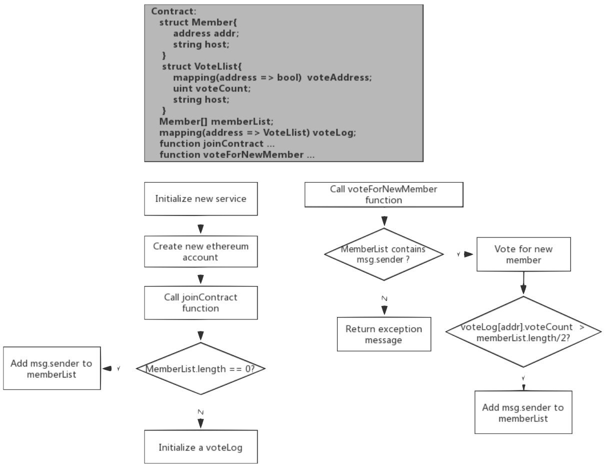
Blockchain is the carrier that actually stores and persists database logs, and smart contracts are the entry point for storing logs. The core of its design is to strictly guarantee the following requirements when storing data:

Management of contract access users. Regardless of the consensus phase of the incremental log between nodes or the subsequent log information on-chain phase, its essence is to ensure Byzantine fault tolerance through a voting pair mechanism, then the system needs to be able to obtain a set of secure and reliable user addresses and nodes Information is stored in smart contracts, which is a very good choice. In addition, smart nodes can also be used to complete user node management. For example, when a new node accesses, a smart contract can be used to ensure that most nodes have access to the new access node. Consensus.

Complete log data storage. Must contain core information such as executable commands, command initiators, command initiation timestamps, log sequence numbers, and so on.

Security for inserting log information. Smart contracts need to perform a security check on the initiator of the log insertion operation. In addition, it is more important to verify whether the log information is the result of consensus of most nodes.

Contract access address management



```
Contract:
  struct Member{
      address addr;
      string host;
   }
   struct VoteLlist{
      mapping(address => bool)  voteAddress;
      uint voteCount;
      string host;
   }
  Member[] memberList;
  mapping(address => VoteLlist) voteLog;
  function joinContract ...
  function voteForNewMember ...
```

Initialize new service
Create new ethereum account
Call joinContract function
MemberList.length == 0?
Y
Add msg.sender to memberList
N
Initialize a voteLog
Call voteForNewMember function
MemberList contains msg.sender ?
Y
Vote for new member
N
Return exception message
voteLog[addr].voteCount > memberList.length/2?
Y
Add msg.sender to memberList

Figure 6: CyberVein blockchain smart contract address management process.

Each database management smart contract maintains a member list, and maintains several Cybervein account addresses and node addresses corresponding to the addresses in the list. Only accounts in this list can modify log data in the contract. After a node service node is started, the service program will create a Cybervein account, which uniquely corresponds to the current service program instance. The service program will initiate an application to join the member list. Each node in the database determines that the application is "passed" or “reject" after voting.

Database log management

Log information is stored in the contract as an array. Stored in data in a specific data structure. After users vote through nodes, users can call functions in the contract or access log information in smart contracts. The data structure includes the executable command of the log, the command initiator, the command initiation time stamp, the log sequence number, and other information. At the same time, the contract also maintains the currently confirmed log number. In addition to log information, the contract also contains necessary elements such as a signature list, a judgment member function, a member list, and a verification function.

After the database operation reaches a consensus among the nodes, each node will call the smart contract to write the operation log to the blockchain.

Consensus mechanism of incremental operation log based on PBFT algorithm

For the above storage structure, to ensure that the data of non-Byzantine nodes in a decentralized database system is consistent, it is necessary to ensure the consistency of the base db contract and the operation log. The base db contract is used as the data base during the database operation The frequency of change is low, and the operation log is continuously increasing with the operation of the database, which requires high consistency, real-time and other aspects. A consensus mechanism is needed to ensure the synchronization of the incremental logs. Since the data in the database is not changed, synchronization between nodes is not required, and it can be returned directly through the local database without the need to modify the data on the blockchain or initiate operations such as voting. The Cybervein database system uses the PBFT algorithm to make the basic consensus algorithm, and some changes are made to its algorithm to adapt the database operation process to achieve the consistency and consensus of the database at the operational level.

Client requests distribution

In the PBFT algorithm used by the Cybervein database system, a Byzantine fault-tolerant master node exists in each database, which can ensure that the entire system still runs normally when the master node is the wrong node. The database client maintains a view ID locally, and the Byzantine fault tolerance of the master node is controlled by the view. Each time the client performs an operation, it will receive a response with a view ID to update the local view ID. However, the ID is not necessarily correct, because in a multi-user scenario, operations between different clients cannot be sensed, so in a concurrent situation, the view ID maintained by the client may not be correct. However, when requesting a service, the client will still use this view ID to calculate, according to:

$$p = v \times \mod \left|R\right| (2)$$

Calculate the master node, where $\left|R\right|$ is the total number of nodes, and $\left|R\right| = 3f + 1$ which f is the maximum number of Byzantine nodes in the system. The client requests a node that it considers to be the master node. Request information:

Among them o are the specific requested operation instructions, t is the current time stamp of the requesting node, *addr* is the address of the requesting node account, and *Sign* is the signature of the blockchain account.

When the single-node Database service receives a write request from a client, if it is the master node, it enters a three-phase submission, and if it is not the master node, it forwards the request to the master node.

Three-phase commit

When the master node receives a request from the client, it will officially begin the three-phase submission process.

After receiving the REQUEST request, the master node will generate a global incremental log sequence number n, which is generated by adding one to the sequence number currently maintained by the master node. Then the master node will attach the current view ID and generate a summary d of the REQUEST. Attach the signature of the master node, generate a PRE-PREPARE request, and broadcast all nodes including the master node, where m is the REQUEST message and is a signature $Sign_p$ of the three messages v, n, and d:

$$< PREPARE < v, n, d > , m, addr_p, Sign_p > (3)$$

After all nodes receive the pre-prepare request, the request that passes a series of verification conditions will be accepted by the node, and when the node accepts the PRE-PREPARE request, the node state will be converted to the PRE-PREPARED state.

When the node enters the PRE-PREPARED state, it will broadcast a PREPARE request with the content of the account address and signature $addr_i, Sign_i$ of the current node:

$$< PREPARE < v, n, d > , addr_i, Sign_i > (4)$$

At the same time, the node will collect PREPARE requests from other nodes. After receiving PREPARE requests from other nodes, it will verify that the current v, n, d of this node is equal to the request, and verify that the signature of the requesting node is correct.

When the number of correct PREPARE requests received from different nodes is greater than or equal to 2f, the node enters the PREPARED state.

When the node enters the PREPARED state, it will broadcast a COMMIT request, requesting content:

$$< COMMIT < v, n, d > , addr_i, Sign_i > (5)$$

In the PREPARE phase, nodes will collect COMMIT requests from other nodes at the same time, and verify the correctness of each request. When receiving COMMIT requests from different nodes greater than or equal to 2f, the node enters the COMMITTED state, and then the local service calls the database Service, execute the execution instruction of REQUEST. And the information about the timestamp, instruction, serial number and execution result are stored in the service program, and when the service program receives the same request again, it can return directly. After the local database service executes the request, it will generate a result r, and the node will return the result r after encapsulation, attach a signature, and return it to the caller:

After consensus:

$$< REPLY < v, t, addr_r eq, r > , addr_i, Sign_i > (6)$$

At the same time, the log sequence number of the successful execution is recorded locally. In the end, all signed return information will be verified by the client. When the exceeding nodes pass $f + 1$

verification of the same result, the client obtains the final execution result.

Log asynchronous write to smart contract

At the end of the above three-phase submission, the logs that have been agreed and executed by the nodes need to be synchronized to the smart contract of the blockchain. When replying to the REPLY message, an UPLOAD request will be broadcast:

$< UPLOAD < n, t, o > , addr_i, Sign_i >$ (7)

Each node will receive UPLOAD requests from other nodes. After receiving the UPLOAD request, it will first call the smart contract's query interface to determine whether the log of the serial number already exists in the contract. If it exists, the request will not be processed. If the log does not exist, collect and save these requests. If for a certain serial number n and the number of requests with the same content exceeds $f + 1$, the smart contract function is called and the log is inserted into the contract. After the signature is verified by the smart contract function, it is written into the contract.

Application Layer

Artificial intelligence has been a hot word in the last year or two. In the areas of image classification, speech recognition, text analysis, computer vision, natural language processing, and autonomous driving, a large number of artificial intelligence and machine learning models are indeed making our lives easier and faster. Technically speaking, most of the current AIs are actually based on statistics and some machine learning methods are playing a role. The core of machine learning is to emphasize that algorithms can automatically learn models based on given data. So far, this solution works perfectly. As long as you have sufficient permissions to access the data, it is almost foreseeable that in the near future, we will fully realize AI. However, countries around the world have begun to attach importance to data privacy security protection regulations. The promulgation of the cryptography law also marks China's determination to build a harmonious network society and seize the opportunity in the fourth industrial revolution in the future. Passwords can be classified according to different criteria. According to the type of protection information, passwords can be divided into: core passwords, ordinary passwords, and commercial passwords. The classification of the security level of passwords means different levels of data security, which means that different types and levels of data will become more and more stringent in the review of future interactions.

The law does not allow us to do crude data aggregation. So what can we do legally to solve this data silo problem? This question should be enough to cause deep thinking for artificial intelligence scholars and practitioners, because it is likely that this dilemma is the cause of the next artificial intelligence winter. Therefore, CyberVein is committed to solving the problem of data silos and using federal learning on the CyberVein architecture.

The data of all parties are kept locally, without revealing privacy or violating regulations;

A system where multiple participants combine data to build a virtual common model and benefit from it together;

Under the federal learning system, the identities and status of each participant are the same;

The modeling effect of federated learning is the same as the effect of modeling the entire data set in one place, or there is little difference (under the conditions of user alignment or feature alignment of each data);

In the case of misaligned users or features, transfer learning can also achieve the effect of knowledge transfer by exchanging encryption parameters between data. Federal learning enables two or more parties to use the data to cooperate with each other without the data being local, thus solving the problem of isolated databases.

CyberVein-based federal learning

To further clarify the idea of federal learning, we define it as follows:

When multiple data owners (such as enterprises) $F_i, (i = 1...N)$ want to combine their respective data D_i to train a machine learning model, the traditional approach is to integrate the data to one party and use the data $D = D_i, (i = 1...N)$ to train and obtain the model M_{sum}. However, this solution is often difficult to implement due to legal issues such as privacy and data security. To address this, we propose federal learning. Federal learning means that these data owners F_i can perform model training and obtain the calculation process of the model M_FED without giving their own data D_i, and can ensure that the gap between the effect V_{FED} of the model M_{FED} and the effect V_{sum} of the model M_{sum} is small enough, that is:

$$\left|V_{FED} - V_{SUM}\right| < \delta \ (7)$$

Here δ is an arbitrarily small positive value.

Considering that there are multiple data owners, the data set D_i held by each data owner can be represented by a matrix. Each row of the matrix represents a user, and each column represents a user characteristic. At the same time, some data sets may also contain labeled data. If you want to build predictive models for user behavior, you must have labeled data. We can call user features X and label features Y. For example, in the financial field, the user's credit is the label that needs to be predicted; in the marketing field, the label is the user's purchase wish Y; in the education field, it is the degree of knowledge of the student. User features X are labeled to form the complete training data (X, Y) . However, in reality, it is often the case that the users of different data sets are not identical, or the user characteristics are not identical. Specifically, taking federal learning with two data owners as an example, the data distribution can be divided into the following three cases:

The user feature $(X_1, X_2, \ldots)$ overlap of the two datasets is larger, while the user $(U_1, U_2 \ldots)$ overlap is smaller;

The user $(U_1, U_2 \ldots)$ overlap of the two datasets is larger, while the user feature $(X_1, X_2, \ldots)$ overlap is smaller;

The overlap between user $(X_1, X_2, \ldots)$ and user feature $(U_1, U_2 \ldots)$ in both datasets are small.

In order to deal with the above three types of data distribution, we divide federal learning into horizontal federal learning, vertical federal learning, and federal transfer learning (see Figure 7).

We take the scenario of two data owners (ie, enterprises A and B) as an example to introduce the system architecture of federated learning, which can be extended to scenarios with multiple data owners. Suppose companies A and B

If we wish to jointly train a machine learning model, and their business systems each have relevant data for their users. In addition, Enterprise B has label data that the model needs to predict. For data privacy and security reasons, A and B cannot directly exchange data. At this point, a model can be built using the federal learning system, and the system architecture consists of two parts, as shown in Figure 8.

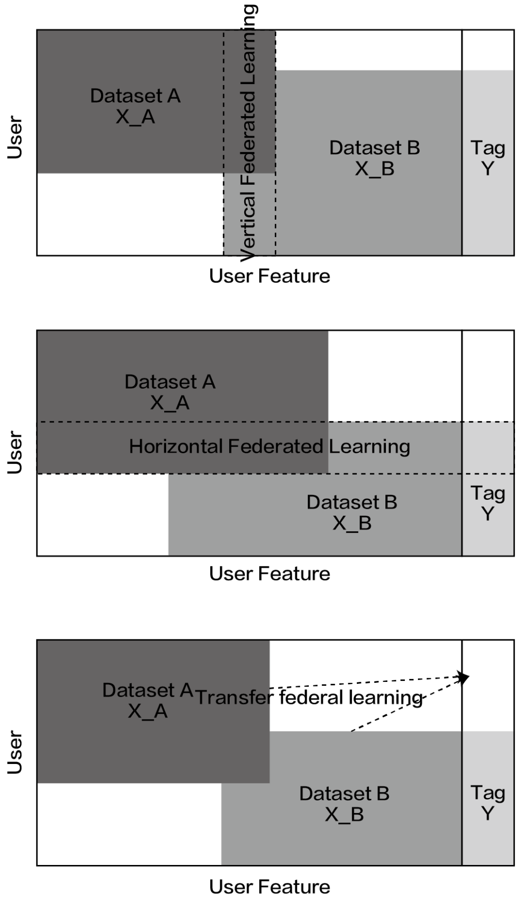


Figure 7: Standards for horizontal, vertical, and transfer learning in federal learning

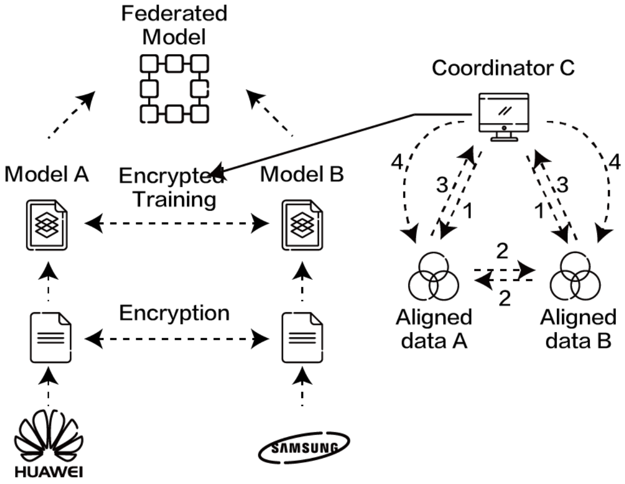


Figure 8: Federal learning federation model building process.

Part I: Alignment of encrypted samples. Because the user groups of the two companies are not completely coincident, the system uses encryption-based user sample alignment technology to confirm the common users of both parties without revealing their respective data, and does not expose users who do not overlap each other. In order to model the characteristics of these users.

Part II: encryption model training. After identifying a common user population, you can use this data to train machine learning models. In order to ensure the confidentiality of the data during the training process, a third-party collaborator C needs to be used for encryption training. Taking the linear regression model as an example, the training process can be divided into the following 4 steps (as shown in Figure 8):

Step 1: Collaborator C distributes the public key to A and B to encrypt the data that needs to be exchanged during the training process;

Step 2: Interaction between A and B in encrypted form used to calculate the intermediate result of the gradient;

Step 3: A and B are calculated based on the encrypted gradient value, meanwhile, B calculates the loss according to its label data, and summarizes these results to C. C calculates the total gradient by decrypting the result.

Step 4: C sends the decrypted gradient back to A and B respectively; A and B update the parameters of their respective models according to the gradient.

Iterate the above steps until the loss function converges, thus completing the entire training process. In the process of sample alignment and model training, the respective data of A and B are kept locally, and the data interaction during training will not cause data privacy leakage. Therefore, the two sides can realize the cooperative training model with the help of federal learning.

Part III: Effect motivation. A major feature of federated learning is that it solves the problem of why different institutions should join the federated common modeling, that is, after the model is built, the effect of the model will be shown in practical applications and recorded on the CyberVein DAG. Institutions with more data will see better effects of the model. The consensus of the PoC contribution value is reflected in their own contributions and contributions to others. The effectiveness of these models on others is federated to individual agencies on the federal mechanism and continues to inspire more agencies to join this data federation.

The implementation of the above three steps has considered the privacy protection and effects jointly modeled among multiple institutions, and has considered how to reward the institutions that contribute more data, which is realized by the PoC consensus mechanism. Therefore, federal learning is a “closed loop” learning mechanism.

Cross-chain technology

The current blockchain technology has many bottlenecks in characteristics such as performance, capacity, privacy, isolation, scalability, etc., and it cannot be applied to commercial applications on a large scale in the short term. In addition to the high technical barriers and limited performance of the blockchain itself, there is also an important factor. At present, each individual blockchain network is a relatively independent network. Data and information cannot be interconnected and there are isolated islands of information. The problem. Collaboration between different blockchain networks is difficult, which greatly limits the development of blockchain applications. Therefore, "cross-chain" has become an important issue of concern for the development of blockchain. The significance of "cross-chain" is to achieve interconnection and interoperability, so that various chain-like structures may be intertwined, so as to better realize its own network value.

CyberVein took the lead in implementing cross-chain technology and multi-chain fusion, and its main goals are:

Mutual conversion of assets: Meeting the exchange needs of different digital currency assets can build a more efficient and low-cost digital economic network.

Satisfy the atomicity of transaction transactions: that is, there will be no interruption in the transaction transaction, or all execution will be completed, or they will be interrupted at the same time.

Oracle problem: Implement mutual awareness between chains and ensure that information can be read from each other. For example, ETH and CVT belong to two independent and autonomous systems. By implementing cross-chain technology, information can be communicated and other chain's dynamics can be received in time.

Asset collateral: read and complete cross-chain information of smart contracts; if two independent chains can achieve the same smart contract, cross-chain must be used. For example, the DAG architecture itself is a mesh structure. It is possible for the DAG and the blockchain to complete the chain and chain event transmission and start the smart contract.

Industrial application

Artificial Intelligence

Many industries are affected by factors such as intellectual property rights, privacy protection and data security, the data that AI needs to obtain cannot be directly aggregated for training in machine learning models. The CyberVein network architecture is highly similar to the neural network model in artificial intelligence, and the artificial intelligence uses federal learning for data training, which fully protects user privacy and data security, without exporting corporate or personal data, building a machine learning big data analysis and model for the third parties which will have broad application prospects in the sales, finance and supply chain industries. CyberVein uses the characteristics of federal learning to break through the barriers of system dynamics and heterogeneous equipment, building a unified characteristics space, and conduct multi-user and multi-device collaborative training based on the unified characteristics space, bringing users a unified, continuous and personalized service experience, thus achieving the common benefits of multiple parties. At the same time, CyberVein also borrows the idea of transfer learning to solve the problem of heterogeneity of user and user characteristic data, mining common knowledge between data and using it to break through the limitations of traditional artificial intelligence technology.

Smart Cities

How to solve the problem of sharing a large amount of smart city data, construction funds for smart city systems duplication, and providing more secure data to more enterprises are all issues that smart city construction has to face. City managers hope to build their own "data centers" efficiently and quickly, and share data and data analysis with "data centers" in other cities to promote sustainable urban development. Through CyberVein, data island bridges can be used to achieve data interconnection between different institutions. CyberVein provides city administrators with a comprehensive platform that can program and manage data storage, data interoperability and data access rights, etc., providing people with smarter, more convenient and better public services in fields such as education, employment, retirement, precision poverty alleviation, medical health, anti-counterfeiting, food safety, public welfare and social assistance. Government agencies at all levels, large enterprises and institutions, internet companies and other institutions can use data island bridges to ensure cross-domain modeling under the premises of security and credibility, lossless accuracy, diverse scenarios, ease of use, lightweight deployment, credible profit division, legal compliance.

The Smart Finance

The policy points out that it is necessary to seize the opportunities of blockchain technology integration, function expansion and industry segmentation, and to use blockchain to promote data sharing, optimization of business processes, reduction of operating costs, improvement of collaboration efficiency, and construction of a trusted system. It is necessary to promote the deep integration of blockchain and the real economy, and solve the problems of SME loan financing difficulties, bank risk control difficulties, and department supervision difficulties. CyberVein establishes a smart financial system, and applies emerging technologies such as the Internet of Things, big data, machine learning and blockchain to the solution of smart payment, smart financial management, smart banking, smart insurance, smart securities and even smart risk control, becoming the underlying technology for smart finance. For example, to support small and micro enterprise financing loans, loan-qualified internet financing platforms, small loan companies, banks, etc., they can model the data locally through federal learning, and the participating institutions share the final risk control and prediction models for loan issuance. Ensuring data security, and data will not be exported. Improve forecasting capabilities and share model effects, providing the underlying guarantee of financial infrastructure for financial information.

The Internet of Things

The Internet of Things has a wide range of applications in various fields. The number of different types of device connections and data transmission will reach unprecedented heights, making the data stored on the central server facing huge tests such as data transmission, storage cost, access efficiency, performance stability and data security. In CyberVein's network, there is no central server that manages data, reducing the pressure on the traditional central computing of Internet of Things, and data is no longer controlled by the central server. The transmitted data is encrypted, and the data will not be tampered with and lost arbitrarily, eliminating trust issues such as security. IoT service providers can share resources on the CyberVein chain, and users can directly settle between various operators across the entire network, reduce communication and transmission costs, and realizing value interconnection.

Database

At present, organizations in various industries have huge data sharing and complementary needs. However, due to the lack of economic stimulus and technical means, most of these data exist in independent and separate databases, which cannot be supplemented and shared with sufficient value, resulting their economic and social potential not being applied. CyberVein, as a distributed database system that realizes data value definition and data management, enables easy data interaction and management between businesses. Its consensus mechanism and encryption improve information transparency, ensuring highly secure information and data sharing, and allowing related companies to quickly establish trust, providing a management platform that can quantify the value of data, linking data demand and data suppliers in terms of data storage and data rights management etc. This database network can also be used independently, providing anytime connection, interaction, and also supports multiple parties to store, query, analyze, mine, and utilize data simultaneously online; data throughput and transaction information confirmation efficiency is faster and better than traditional networks, and stored information is traceable at any time and cannot be tampered with.

Project Outlook

CyberVein has a returnee team composed of high-level, highly educated talents. The core members are from internationally renowned universities such as Imperial College of Technology, Carnegie Mellon University, Cambridge University, and have rich experience in blockchain development. In 2018, CyberVein and Zhejiang University jointly established the "Zhejiang University-Shumai Chain Joint R & D Center" to jointly build a blockchain technology R & D system. They also published a professional academic on the concept of PoC in the science and technology special issue of BIG SDM Scientific Data Conference Paper "Data quality transaction on different distributed". In addition, combined with practical experience and research directions, he also published several articles such as "Exploratory Analysis for Big Social Data Using Deep Network", "An artificial intelligence based data-driven approach for design ideation", and "Blockchain-Based platform for Distribution AI". research paper.

In specific practice, CyberVein has been analyzing the problems encountered in the existing blockchain design process and proposing solutions. Has provided a blockchain tax solution for Taishi tax rebate, through the alliance chain method, the construction of electronic invoices and tax blockchain transfer and integration application platform; assisted Zhixing LAB in technical development and technical problems, providing them with more professional Blockchain database; provides smart city solutions for the governments of North Jiangsu, Zhoukou City, and Henan Province. It will also extend the solution overseas to provide blockchain smart retail services for Australian e-commerce; provide digital building solutions for Australian real estate AEC; provide US marketing groups with digital anti-fraud technology in the field of digital marketing; and CS Pay Capital Pay provides industry technical services such as smart financial solutions.

At the same time, CyberVein, as a public and cross-chain open to the world based on the world pattern, always pays attention to the development status and trends of blockchain technology, so as to improve its ability to use and manage blockchain technology. It will be widely implemented in the real economy within 3-5 years, accelerate the process of "trusted digitalization", drive financial "deadliness" and serve the real economy, and play a role in building a network power, developing a digital economy, and helping economic and social development.

References

[1] H. Abusalah, J. Alwen, B. Cohen, D. Khilko, K. Pietrzak, and L. Reyzin. Beyond Hellman's time-memory trade-os with applications to proofs of space. In ASIACRYPT (2), volume 10625 of LNCS, pages 357{379. Springer, 2017.

[2] S. Dziembowski, S. Faust, V. Kolmogorov, and K. Pietrzak. Proofs of space. In R. Gennaro and M. Robshaw, editors, CRYPTO 2015, volume 9216 of LNCS, pages 585–605. Springer, 2015.

[3] Chia Network. https://chia:network/, 2017.

[4] G. Ateniese, I. Bonacina, A. Faonio, and N. Galesi. Proofs of space: When space is of the essence. In M. Abdalla and R. D. Prisco, editors, SCN 14, volume 8642 of LNCS, pages 538{557. Springer, Heidelberg, Sept. 2014.

[5] Ari Juels and Burton S. Kaliski, PORs: Proofs of Retrievability for Large Files, CCS '07 Proceedings of the 14th ACM conference on Computer and communications security, 978-1-59593-703-2,USA

[6] Hovav Shacham1 and Brent Waters, Compact Proofs of Retrievability, J. Pieprzyk (Ed.): ASIACRYPT 2008, LNCS 5350, pp. 90–107, 2008 International Association for Cryptologic Research 2008.

[7] Kevin D. Bowers, Ari Juels, Alina Oprea, HAIL: A High- Availability and Integrity Layer for Cloud Storage, ACM 978-1- 60558-784-4 /09/11, USA.

[8] Fecher, B., Friesike, S., Hebing, M., Linek, S., Sauermann, A.: A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing (2015)

[9] Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137–144 (2015)

[10] Giannakis, M., Louis, M.: A multi-agent based system with big data processing for enhanced supply chain agility. Journal of Enterprise Information Management 29(5), 706–727 (2016)

[11] Sravan Kumar R, Ashutosh Saxena, Data Integrity Proofs in Cloud Storage, 978-1-4244-8953-4/11/@ 2011 IEEE.

[12] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)

[13] Hassabis, D.: Artificial intelligence: Chessmatch of the century. Nature 544(7651), 413 (2017)

[14] Kevin D. Bowers, Ari Juels, Alina Oprea, Proofs of Retrievability: Theory and Implementation, CCSW'09, Journal of Systems and Software, v.85 n.5, p.1083-1095, May, 2012.

[15] John Walker, S.: Big data: A revolution that will transform how we live, work, and think (2014)

[17] Leiding, B., Memarmoshrefi, P., Hogrefe, D.: Self-Managed and Blockchain-Based Vehicular Ad-Hoc Networks. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. pp. 137–140. ACM (2016)

[18] Lord, N.: The Third Party Data Breach Problem. URL: https://digitalguardian.com/blog/third-party-data-breach-problem (2017), (Accessed April 23, 2018)

[19] Mardis, E.R.: Dna sequencing technologies: 2006–2016. Nature protocols 12(2), 213 (2017)

[20] Marr, B.: How Blockchain Will Transform The Supply Chain And Logistics Industry.URL: https://www.forbes.com/sites/bernardmarr/2018/03/23/how-blockchain-will-transform-the-supply-chain-and-logistics-industry/#335ee6a85fec (2018), (Accessed April 23, 2018)

[21] Norall, S.: 3PL vs 4PL: What are these PLs, Anyway? Layers of Logistics Explained. URL:http://cerasis.com/2013/08/08/3pl-vs-4pl/ (2013), (Accessed April 23, 2018)

[22] L. Ren and S. Devadas. Proof of space from stacked expanders. In M. Hirt and A. D. Smith, editors, TCC 2016-B, Part I, volume 9985 of LNCS, pages 262–285. Springer, Heidel- berg, Oct. / Nov. 2016.

[23] M. Rosenfeld. Analysis of hashrate-based double spending. CoRR, abs/1402.2009, 2014.

[24] Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, (2016)

[25] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre,L., Kumaran, D., Graepel, T., et al.: Mastering chess and shogi by self-play with a

general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815 (2017)