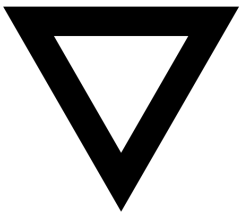


Whitepaper

{Nick Zhu}@tethys.ai
Last Updated: May 7, 2018
Version 1.7

Abstract

This whitepaper introduces a novel decentralized semantic information collection network - the Tethys Network. It is designed to overcome the limitations in the centralized web crawling methods deployed by search engines today. Tethys's decentralized crawling architecture leverages a unique blend of blockchain, artificial intelligence, and human intelligence to solve the throughput, temporal, and depth of extraction challenges that the centralized crawling architecture is facing today. With this in mind, the Tethys Network is designed to index and record captured semantic information in a temporal state transition format, using a special purpose, high-throughput blockchain implementation. Finally, the network will supply a native currency, the Tethys Token, which will become the sole currency used in its ecosystem for contract payment and to incentivize end-users' participation.

The Tethys Network introduces four fundamental concepts and innovations:

- The Tethys blockchain is a permissioned partitioned blockchain implementation utilizing Practical Byzantine Fault Tolerance (PBFT) to achieve very high transaction throughput required for web-scale information collection.
- A Proof-of-Reputation based, decentralized information mining network running on end-user's devices that is resilient to DDoS and Sybil attacks.
- Pre-built deep learning models designed and trained to make classification, recognition, and extraction of semantic information in both textual and image form relatively easy, with unstructured noisy web content.
- A statistical consensus model that is capable of establishing statistical truth through inductive reasoning for information collected, thus rewarding users accordingly.

The rest of this paper will describe these concepts and the Tethys Network in more detail.

TABLE OF CONTENTS

Introduction

The Semantic Web is a Web of actionable information derived from data, with “meaning” through interpretation by a semantic system. In the original 2001 Semantic Web article published on Scientific American, Tim Berners-Lee, the inventor of the World Wide Web, and his co-authors James Hendler and Ora Lassila, declared that “The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.”[1] More than a decade and a half later, we argue that today’s Web is still a far cry from this original vision.

Early attempts on the semantic web have focused primarily on tagging through eXtensible Markup Language (XML), Resource Description Framework (RDF), and more recently with W3C Microdata standard, however, tagging on web-scale has proven to be challenging in practice. Christian Bizer and his team have observed in their quantitative analysis on semantic web that microformat data, excluding organizational contact information, has less than 5% coverage.[2] More recent research has shown promising new paradigms in semantic information retrieval with both noisy web textual content[3] as well as imagery content[4]. As similarly stated in Semantic Web Revisited[5], we also believe the future of the Semantic Web is in developing and understand distributed information systems, systems of humans and machines, operating on a global scale with AI playing a central contributing role.

On the other hand, since its inception blockchain technology has been widely seen as transformative in its most natural application - value transfer using decentralized digital currency. However, it has incredible transformative power in information science as well. A general purpose blockchain data structure that is capable of recording arbitrary chronological state transition is an excellent tamper resistant choice for providing a long-term memory for the World Wide Web. This ability to record and recall information with point-in-time accuracy is a crucial component of a Semantic Web since a large percentage of semantic information are time sensitive for example product stock level and hotel price.

In this paper, we propose a novel information collection architecture combining a unique blend of decentralization, blockchain, and both machines and humans’ intelligence to improve significantly over the centralized shallow web crawling method employed by the search engines today.

---

[1] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." *Scientific american* 284.5 (2001): 34-43.
[2] Bizer, Christian, et al. "Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis." *International Semantic Web Conference*. Springer, Berlin, Heidelberg, 2013.
[3] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." *Advances in neural information processing systems*. 2015.
[4] Huang, Changqin, et al. "Large-scale semantic web image retrieval using bimodal deep learning techniques." *Information Sciences* 430 (2018): 331-348.
[5] Shadbolt, Nigel, Tim Berners-Lee, and Wendy Hall. "The semantic web revisited." *IEEE intelligent systems* 21.3 (2006): 96-101.

Problems

To accurately assess and evaluate the motivation behind the creation of Tethys, a thorough study of the issues and shortcoming of the current generation centralized web crawling architecture is required. The following diagram illustrates how web crawling is conducted in a centralized architecture.

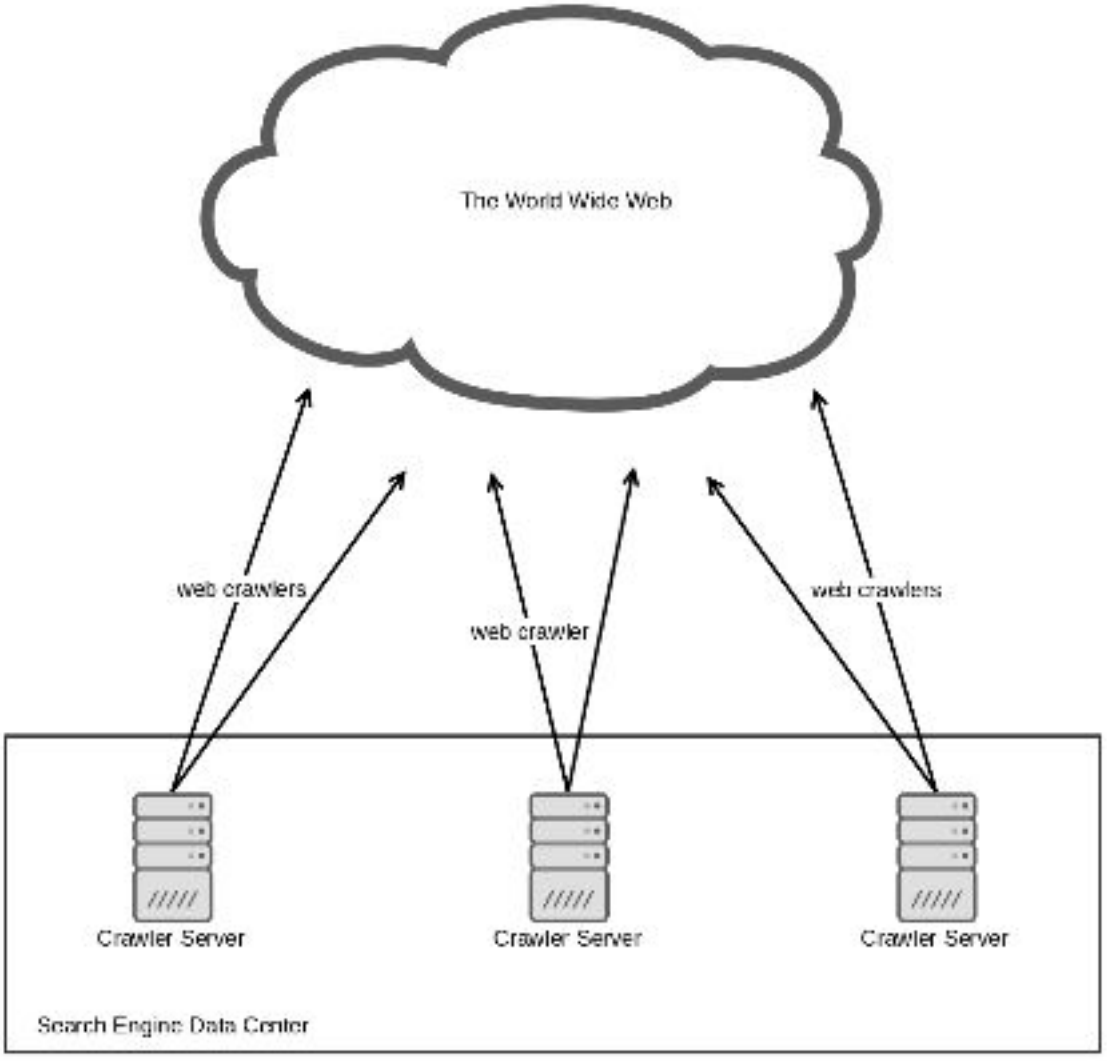


Figure 1: *Centralized web crawling architecture deployed by the search engines*

The search engines send out web crawlers, specialized bot agent programmed to systematically visit the World Wide Web, from servers operated within their data centers. Obviously, this illustration is much simplified, a large search engine will deploy dozens or more types of specialized web crawlers for different kinds of assets or even different verticals across many geographic regions from multiple data centers, however, the centralized architectural theme remains unchanged. This centralized approach works well with

semi-static and non-semantic information in the shallow web yet facing severe challenges when dealing temporal semantic web information.

Problem 1: Scale

The first challenge might come as a surprise. In spite of the fact that most web users believe that search engines index every single page on the web which is far from being true. The web pages indexed by search engines and accessible directly to anyone using the internet are considered as the Surface Web.[6] The opposite of the Surface Web is the Deep Web which is estimated to contain 500-550 times more information compared to the surface web or roughly accounting for 96% of the total estimated information volume on the web.[7] Other observations have also been made about unexplained fluctuation as high as 50% in reported index size by popular search engines suggesting inclusion or removal of certain content between shallow and deep web.[8] The best way to demonstrate this is to look at a practical example. Let's take the largest online store in the US for example. According to the estimate Amazon.com sells over 562 million of products.[9] At the time of writing, the total number of pages indexed by Google from Amazon.com is around 154 million.[10] In theory, Google's index number should be much larger than the total number of products Amazon sells since every product has its own page and on top of that Amazon.com also has numerous non-product pages such as FAQ, Amazon Jobs, Amazon Prime, and etc. What is causing this obvious gap? The key issue lies within centralized crawling architecture's inherited rate limitation. Centralized method has a relatively low throughput limit since a search engine has to rate limit their crawler against a particular website before it turns into practically a Denial of Service (DoS) attack. A high rate of crawling will trigger auto defense response ranging from a one-time challenge that requires assistance from a human or a direct ban of the origin IP address.[11] Therefore once the website exceeds a certain size, it becomes impossible for the search engines to index every page. New pages added to these large websites could take days or even weeks before it gets indexed by search engines.

Problem 2: Semantics

The web content of today is a lot more semantic than just merely textual. A number on a particular web page might be a price, discount rate, mortgage rate or payback period for example. The only way to discover what it means requires a comprehensive understanding of the context and the semantic meaning of other critical elements on the page. Current crawling infrastructures lack this ability. Taking a specific product as an example, this is what a search engine displays on its results page:

---

[6] Michael K. Bergman. White paper: *The deep web: Surfacing hidden value. Journal of electronic publishing, 7(1), 2001.*

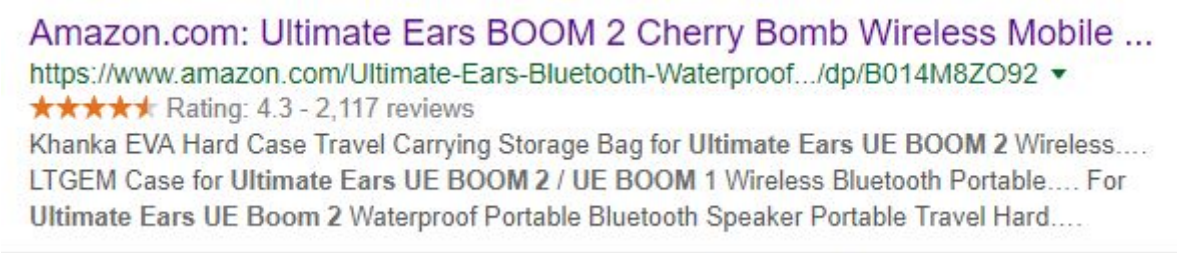
[7] Zhao, Feng, et al. "SmartCrawler: a two-stage crawler for efficiently harvesting deep-web interfaces." *IEEE transactions on services computing* 9.4 (2016): 608-620.

[8] Van den Bosch, Antal, Toine Bogers, and Maurice De Kunder. "Estimating search engine index size variability: a 9-year longitudinal study." *Scientometrics* 107.2 (2016): 839-856.

[9] ScrapeHero. https://www.scrapehero.com/many-products-amazon-sell-january-2018/

[10] Google. https://www.google.ca/search?q=site%3Aamazon.com

[11] Koch, William, and Azer Bestavros. *Reducing web application exposure to automated attacks*. Computer Science Department, Boston University, 2016.

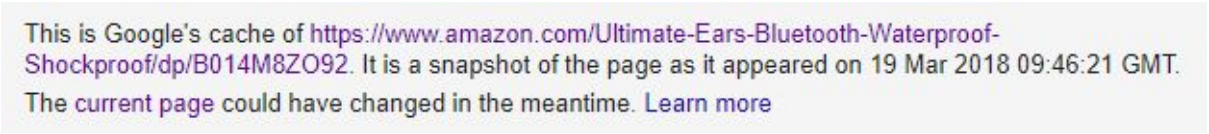


In this example, other than the product ratings, none of the important semantic information of a product (price, discount, color, image, availability, shipping cost, etc,) is available from the search engine. Additionally, if we examine another example, mortgage rate, the result is equally textual centric and lack of semantics.

Web users today were trained to ignore this incomprehensible string of numbers and instinctively knew that the useful information they need is on the bank's website. In other words, users understood this is just an entry point to the deep web where the useful semantic information resides.

Problem 3: Temporal Information

The centralized crawling infrastructure is ideally suited to index static and quasi-static information, e.g. news, articles, and academic publications. Once an academic paper is published it does not change very often nor do newspaper articles; therefore it is perfectly acceptable for search engine only to recrawls these page once a week or less. Nevertheless, for time-sensitive information, this level of crawling frequency becomes unacceptable. For example data on insurance rate, product price, flight ticket, and many other verticals require daily if not more frequent updates to be up-to-date. The following image shows the message returned by a cached version of a page indexed by Google:



This information is more than five days old at the time of writing, and the product price Google indexed was utterly out-of-date from the current selling price on Amazon. This is understood as temporal information retrieval (T-IR) problem in information theory which is one of the key aspects that determine information credibility and usefulness in many verticals.[12] As discussed in Problem 1, if a search engine decides to increase the crawling frequency from weekly to daily or hourly; as a result, it will practically turn crawling into a DoS attack thus triggering website's auto-defense mechanism. Additionally, on the other

[12] Metzger, Miriam. "Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research". *Journal of the American Society for Information Science and Technology*. (2007) 2078–2091.

hand, today's search engines also lack the ability to preserve and retrieve historical information on a subject. This historical information could be invaluable in many verticals for the consumers, for example, the past price history on a particular product or past mortgage rate from a bank. The inability to store and access such history information severely limits the usefulness of the search engines and the web at large.

Problem 4: Information Verification

The final issue, also the greatest one faced by centralized web crawling architecture, is concerning the truthness of the indexed information. In early days, the internet was very host-centric where internet users needed to memorize a large number of useful websites by name or relied on portals and directories to find these sites. Since the advent of search engines, the internet as we know has evolved into an information-centric architecture. An information-centric network has major advantages in large-scale information dissemination.[13] In such network, the value of the network derives from the information it contains within and the search engines play a key role in efficient information dissemination. However, due to the centralized nature of web crawling architecture, it is impossible to establish the truthness of indexed content. What we mean by that statement is to say that the search engines are trying to establish truth by a single observation. Let's revisit the wireless speaker example. The truth about what price Amazon is selling this speaker is only known to Amazon. When web crawler visits the page at a particular time, it faithfully indexes its content as the "ground truth", however, what if at that time the website was malfunctioning or somehow incorrect information was presented to the crawler? A search engine will have no idea and be unable to detect any misinformation in such case. Internet users are trained to place a huge amount of trust on these information middlemen (search engines) despite the fact that they can never verify the truthness of the information. This kind of unreliable "truth" might be no more than an inconvenient annoyance for surface web information since users primarily use them as entry points. In comparison, this kind of flaw is lethal for semantic web information providers since an incorrect price on a product or lending rate for a mortgage will render that kind of semantic information entirely useless. There is no solution to this problem without a complete redesign of today's web crawling architecture.

Solution: The Tethys Network

In this paper, we propose the Tethys Network, the first practical and functionally rich decentralized semantic information collection network that addresses all the issues mentioned above. Tethys consists of the following fundamental concepts and building blocks:

[13] Dannewitz, Christian. "Netinf: An information-centric design for the future internet." *Proc. 3rd GI/ITG KuVS Workshop on The Future Internet*. 2009.

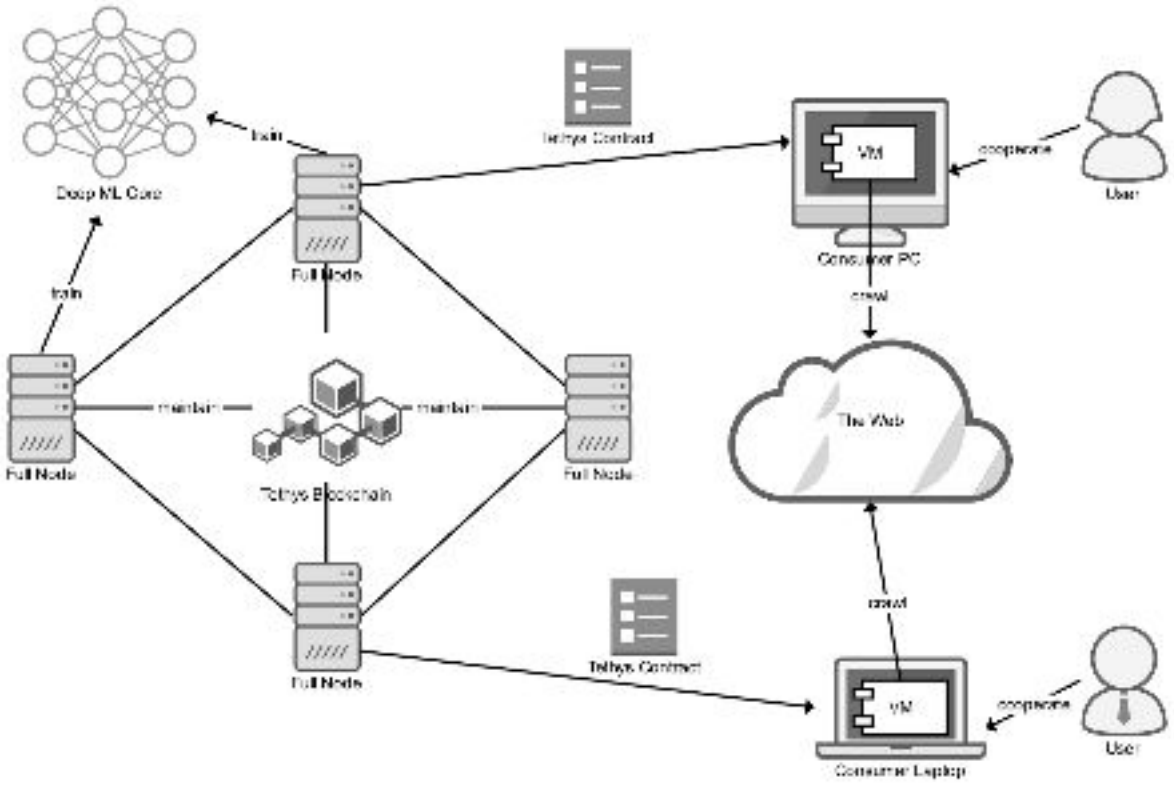


Figure 2: *Tethys Network high-level component overview*

Tethys Blockchain

Tethys blockchain is a special purpose blockchain designed with the following objectives in mind:

- **High Throughput.** Tethys blockchain has to offer scalable throughput to handle information collection and index at web scale. We designed the Tethys blockchain to be a permissioned blockchain meaning full nodes are by invitation with Proof-of-Stake membership. Lightweight client nodes connected to the chain are only capable of submitting queries and transaction requests. Additionally, to achieve consensus in the face of Byzantine failures, Practical Byzantine Fault Tolerance (PBFT) protocol will be implemented to provide fault tolerance. It uses the concept of replicated state machine and voting by replicas for state transition. PBFT based system requires only 3N+1 replicas in order to tolerate N failing nodes with very limited overhead.[14]
- **Temporal Data Structure.** Generalized block storage on Tethys blockchain is extended to not only be able to record transactions in ledger format but arbitrary state transition in chronological order. This generalized storage format enables the blockchain implementation to record time series state transition for any digital asset be it a product sold online or a mortgage offer from a bank.
- **Parallel Scalability.** Permissioned blockchain architecture as mentioned previously alone is not enough to offer unlimited scalability in throughput. Popular blockchain implementations today are not parallel-scalable meaning the throughput of a

[14] Castro, Miguel, and Barbara Liskov. "Practical Byzantine fault tolerance." *OSDI*. Vol. 99. 1999.

particular blockchain is not a function of the number of participating nodes where the throughput is positively correlated to the size of the network. Instead, a majority of the blockchain implementations have constant throughput with a fixed theoretical ceiling, where having more participating nodes results in no increase in throughput regardless whether Proof-of-Work or Proof-of-Stake protocol is implemented. In Tethys blockchain implementation we propose a novel hash partitioned multi-chain architecture where linear parallel scalability can be achieved. This is primarily due to the fact that Tethys blockchain is designed to store chronological state transition instead of transaction ledgers. For financial ledgers, strong sequence and consistent guaranteed is required. For example in a case where A transfers 10 coins to B, then B transfers 10 coins to C these two transactions have to be recorded and verified consistently in the right order regardless who is mining the block or which block eventually gets to be appended to the chain. However, in Tethys' case state transitions can tolerate temporary inconsistency, in other words, we believe temporarily it is acceptable if certain transitions is not immediately consistent or visible from a particular partition. This design choice in favor of an eventual consistency instead of a strong consistency guarantee allows Tethys to partition blocks into multiple chains with a consistent hash function thus achieving unlimited parallel scalability; additionally there are strong evidences shown eventual consistency system often behaves like strong consistency in production.[15]

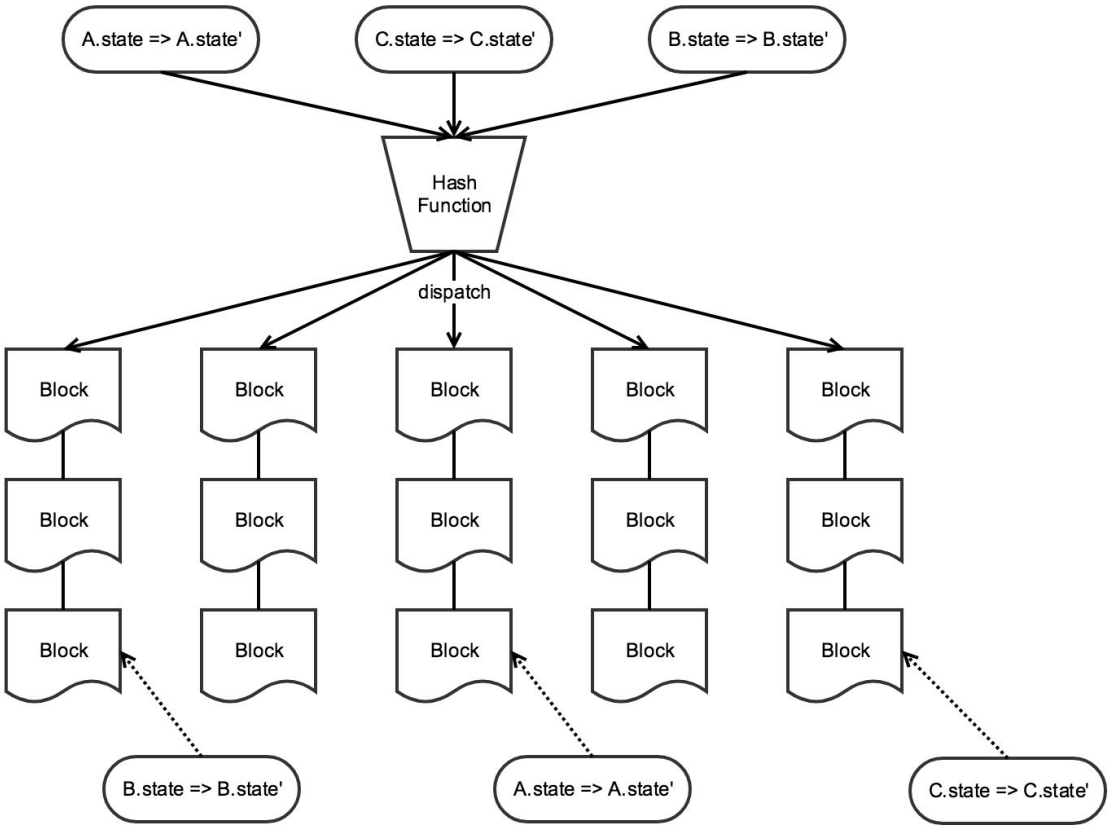


Figure 3: *Tethys parallel multi-chain blockchain architecture*

[15] Bailis, Peter, and Ali Ghodsi. *"Eventual consistency today: Limitations, extensions, and beyond." Queue 11.3 (2013): 20.*

A final note on this is that the Tethys blockchain uses a dedicated unpartitioned chain for recording Tethys token ledgers with strong consistency guarantee required for financial transactions.

- **Indexed States.** Fast lookup time for the current state (world state) is critical for transaction throughput in Tethys. On the other hand, point-in-time query is the foundation for temporal information retrieval (T-IR). To address these two use cases, we chose to introduce an off-chain key-value pair NoSQL database. This database powered approach offers constant O(1) lookup time for world state query and logarithmic O(log(n)) lookup time for point-in-time query.
- **Multi-Channel Support.** For different business clients to leverage the chain for data storage multiple channel structure is required for data segregation and privacy. On Tethys blockchain both public and private channels are supported. Data stored on a public channel can be queried by any client while private channel data is only accessible by the authorized clients.

Deep ML Core

Deep ML Core is a collection of deep machine learning models built and tuned to handle context-aware semantic navigation and information extraction with unstructured noisy web data. The initial release of ML Core will support the following models:

- **Classification Model** is specifically designed to answer a binary question whether a particular page fits a specific type. A sample prediction could answer whether a given page is about a mortgage product or a credit card from a bank.
- **Entity Resolution Model** is designed to leverage syntactic and semantic context information to cluster identical elements together. A typical prediction of this kind of model can answer is whether a given number on a page is about a price of a product for example.
- **Object Detection Model** is a convolutional neural network designed to segment images and classifies regions on a given image. Predictions made by this model typically focuses on recognizing a given image or a region of a given image. For example, this model can answer the question if a given image is about a credit card or a laptop.

Businesses can develop and train sub-models based on the core models with their own data to handle specific niche and vertical. For instance, a business can leverage the classification model to develop a sub-model that identifies whether a web page is about a small business loan product. Additionally, ML Core integration API planned for future releases will also allow businesses to bring their own models to the network so proprietary machine learning models can be leveraged.

Tethys VM

Tethys Virtual Machines (VM) are software agents deployed on end-users' computing device, e.g. desktop, browser, mobile phone, etc. Tethys VM is an execution environment for Tethys Script as well as a lightweight client to the blockchain. A few novel concepts and key components worth highlighting here.

- **Tethys Script** is a specialized programming language designed to automate web crawling tasks. Script interpretation is conducted in a sandbox environment with strong runtime safety policies to make sure the execution is safe on end-user's device. We have also intentionally designed the language to be Turing incomplete with a set of simple and unsurprising constructs for added safety.[16] Finally being Turing incomplete Tethys VM environment is capable of providing guarantees with bounded memory and computation time.
- **Tethys Contracts** are task blueprint written in Tethys Script language. The network supports two different kinds of contracts - automated and manual. An automated contract can be executed by VM without human intervention while a manual contract will require explicit human participation. On the other hand, each contract also has a specific amount of tokens associated as a reward to incentivize end-user participation.
- **Open Standard** is at the core of its design. Tethys VM protocol and specification will be published publicly and can be implemented by any general purpose programming language in various embedded environments for example browser extension, mobile app, desktop software or even directly on a web page. Open source reference implementations will be made available in Phase I.

End User

The last and perhaps most crucial building block in Tethys is its human end-users. End users participation is the most important concept with Tethys which redefines web crawling in general. End-users on Tethys network are the enablers of the following vital functions:

- **Share-Economy.** The end users are the chief enablers of the share-economy facilitated by Tethys network. In essence, end users can place unused computing power, bandwidth as well as their human intelligence on a shared public market in exchange for tokens. Since the computing power and bandwidth are already paid for by the end users thus participating in this share-economy has very low cost. We believe the token based reward can serve as a strong incentive thus foster rapid adoption.
- **Hardware Provider.** Decentralized web crawling hardware in Tethys network are all provided by the end users. In this statement, we are not only talking about the computing device and internet connection but also human intelligence. As we discussed before in the Tethys VM section, there are two different kinds of contracts - manual and automated. Manual contracts are typically designed to collect training data for ML models to solve particular challenges in specific verticals; these contracts require 100% human participation with much higher reward.

[16] Pike, Lee. "Hints for High-Assurance Cyber-Physical System Design." *Cybersecurity Development (SecDev), IEEE*. IEEE, 2016.

Tethys Token Ecosystem

In this section, we will discuss the proposed business ecosystem of the Tethys network; it's primary stakeholders and their respective roles. First of all, let's take a look at the following illustration that depicts the Tethys token ecosystem:

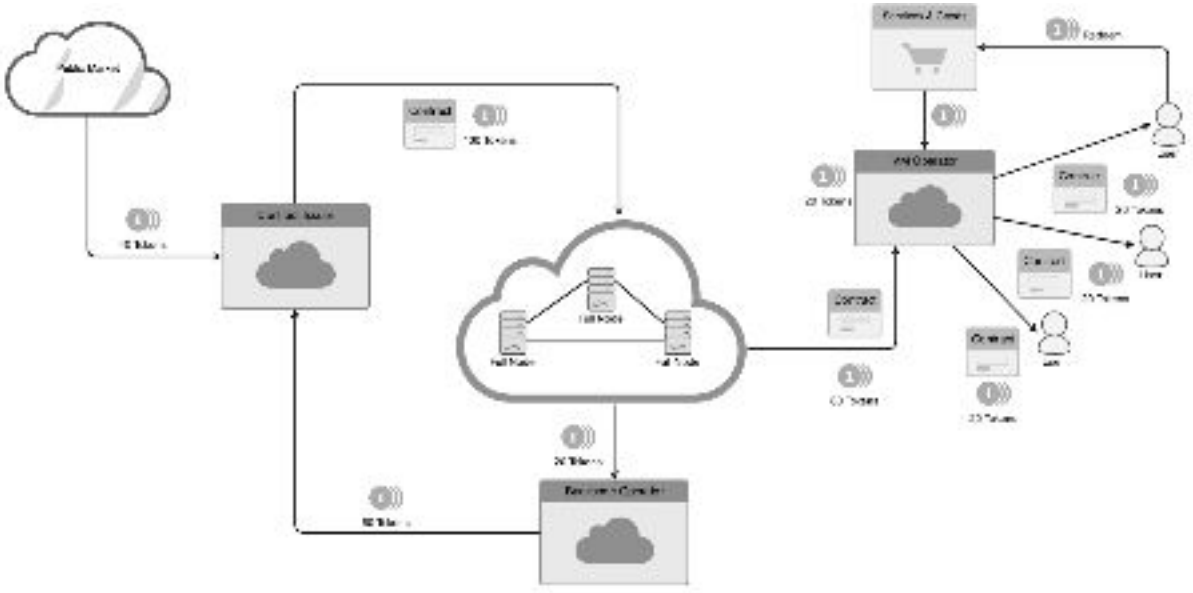


Figure 4: *Tethys token ecosystem*

This illustration demonstrates how contracts and tokens flow through Tethys ecosystem.

Tethys Token

Tethys token is the "currency" used by the network for contract issuance and reward system. These tokens are ERC20-compliant tokens built on Ethereum public chain. Contract issuers need to attach a certain amount of tokens (bounty) with each contract to pay for the service Tethys network provides. All other participating parties get rewarded for every contract fulfilled through the network. The demand for the tokens is strongly and positively correlated to the demand for the service Tethys network provides.

Contract Issuer

Contract issuers are participating businesses who leverage Tethys technology to gain access to semantic deep web information. They perform the following functions in the ecosystem:

- They generate demand for Tethys tokens by purchasing tokens from the market
- They design and issue Tethys contracts for semantic web information collection

Full Node

Full nodes are servers operated by the foundation and authorized partners with special permissions to form Tethys backbone. The primary functions and responsibilities of these nodes are:

- Maintain Tethys blockchain
- Operate a distributed ML core service
- Provide implementation for Tethys consensus protocol for contract fulfillment verification
- Receive 20% of the rewards as a fee for each contract fulfilled

VM Operator

VM operators are participating businesses who offers end-user facing VM implementation embedded in their software environment. They could be mobile application developers, browser extension vendors or even website operators who choose to embed Tethys VM implementation as part of their software offering to their users. These operators provide the following crucial functions in the ecosystem hence are also rewarded for every contract fulfilled through their VM platforms:

- Offer embedded virtual machine environment
- Onboard end-users to Tethys ecosystem
- Receive 20% of the rewards for every contract fulfilled using their edge nodes
- Offer virtual services and goods, in exchange for Tethys tokens, to end users

End User

End users are those consumers who agree to participate in the share-economy created by Tethys ecosystem where a unique combination of computing powers, bandwidth, and optionally human intelligence are required to fulfill received contracts. End users are a crucial part of the system since they are the decentralized edge nodes where semantic web information are collected and analyzed hence a significant portion of the contract bounty goes to reward end users. The principle roles end users, and their hardware play are:

- Supply computing power and bandwidth to execute Tethys contracts
- Optionally offer human intelligence for manual contracts
- Receive 60% of the rewards as bounty for each contract fulfilled
- Can redeem tokens with VM operator for services or goods

Other Key Considerations

In this section, we will cover a few novel concepts and solutions to the unique challenges faced in building a decentralized semantic information collection network.

Tethys Contract Proof of Work

As previously discussed Tethys contracts are essentially programs executed on end-users' devices to collect third party information from the web. In Problem 4: Information Verification section we have also studied the fact that no one other than the owner of such information knows the validity of collected data. This challenge creates a unique problem in establishing Proof-of-Work for Tethys contract. In other words, it is difficult to even answer the question: Did the node perform actual work, and if it did is the result correct? Traditional cryptographic Proof-of-Work algorithm relies on the mathematical asymmetry at the heart of the hash function - finding the right hash value is difficult and time-consuming, in contrary, verifying the hash value is trivial. Lacking this asymmetry when dealing with 3rd party information collection, Tethys contract relies on a consensus-based PoW to verify work performed and establish statistical truth for the information collected. The following diagram illustrates at a high level how statistical truth is established with Tethys consensus layer and quorum system.

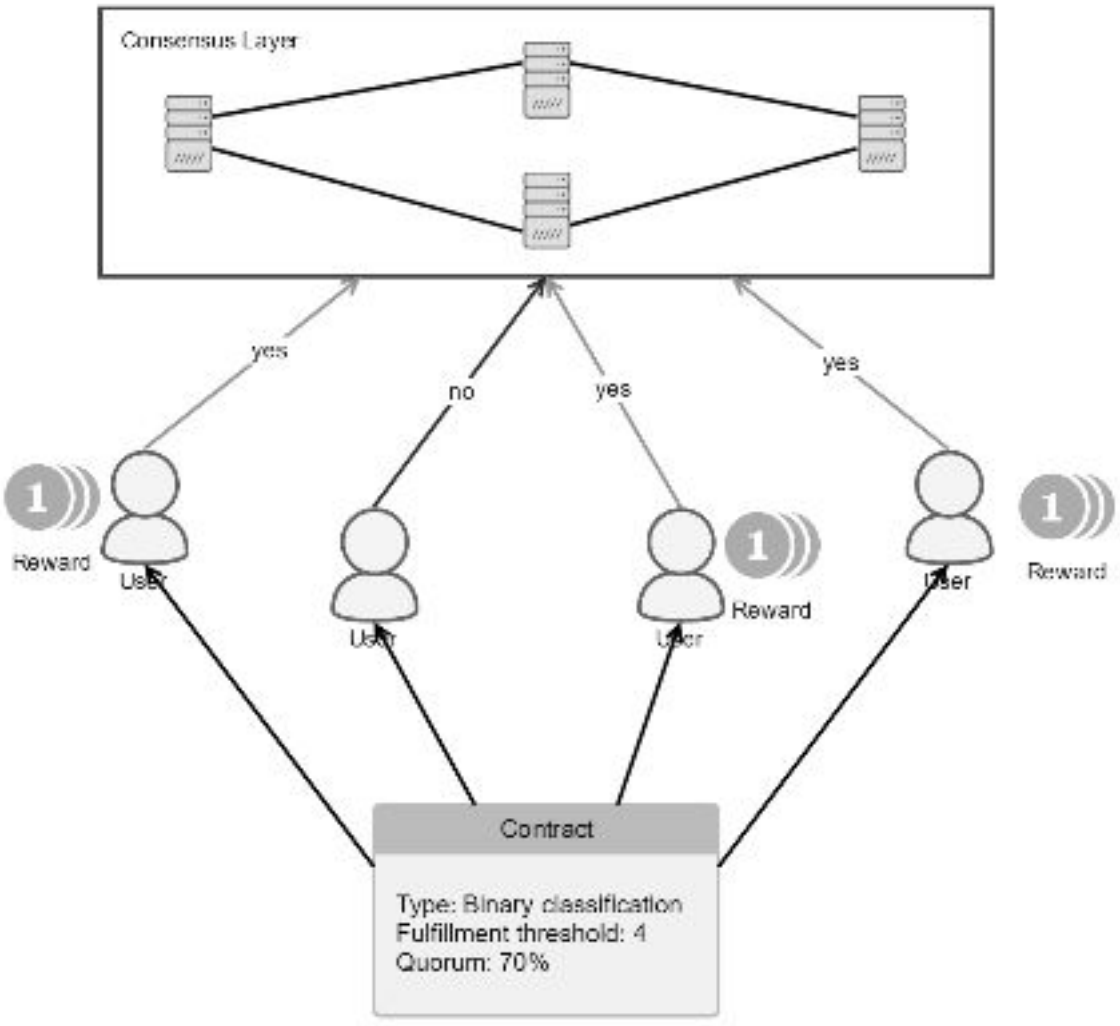


Figure 5: *Tethys Consensus Proof-of-Work*

Let's go over a few essential points worth highlighting in this simple example.

1. **Built-in Tethys Contract Support.** Tethys contract has built-in constructs to allow the issuer to specify the minimum fulfillment threshold meaning at a minimum how many successful executions by edge nodes is required to fulfill this contract. Additionally, the issuer can also specify a quorum threshold, in this example 70%, which will be used by the consensus layer to determine whether statistical truth can be established. In this example, this particular contract is considered successfully fulfilled since four nodes executed the contract and a quorum of 75% has reached consensus. For more complex contracts that generate multiple results, an iterative consensus process is required to go through each result in sequence.
2. **Consensus-based Bounty System.** As illustrated by this example Tethys bounty system is also consensus-based meaning only results that conform to the consensus are counted as valid hence receive the reward. In this example, only those users who answered yes to this classification contract received their rewards.
3. **Cryptographic Checksum.** Each Tethys contract also is embedded with a unique cryptographic puzzle. Every node needs to solve the cryptographic puzzle and append the answer to the contract result before it can submit the result back to the chain. Since verifying the puzzle solution is computationally trivial Tethys consensus layer can perform a preliminary checksum verification before processing the contract result hence protecting the network from DDoS attack.

As discussed in Tethys network, contract issuer can demand a level of confidence in the statistical truth established by the consensus. Higher the fulfillment threshold and quorum size higher the confidence level in established statistical truth, however, this also results in higher execution cost.

Reputation-based Anti-Fraud Protocol

Tethys consensus layer alone is not sufficient in guarding the network against Sybil and 51% attack. Decentralized consensus protocol, if not carefully designed, is typically vulnerable to Sybil attack where attackers attempt to gain 51% of the votes in order to control the consensus process. Sybil attack is specifically designed to overpower consensus system by flooding the protocol with a large number of pseudonymous identities and forged results. Even with cryptographic checksum we introduced in the previous section Tethys consensus protocol is still vulnerable to such an attack; the checksum protection will merely slow down the attack. To defend against this kind of attack, Tethys dedicates a public channel on the blockchain to track and record every participating node's reputation score based on their past performance, for example, number of contracts fulfilled, and number of correct results produced, etc. With this global reputation repository, Tethys backbone can enforce a scaled minimum reputation score for contract fulfillment. Reputation requirement scales positively with the size of the contract bounty, in other words, larger the bounty higher the reputation requirement.

Cold-Start Problem

Minimum reputation requirement for contract fulfillment can effectively deter Sybil attack since new pseudonymous identities created by attackers will not be able to fulfill any contract

thus making the attack less economical. Nevertheless, this threshold also introduces a cold-start problem where newly registered users will not be able to accept any contract. To solve this problem we propose a novel elapsed time-based bootstrap algorithm in which new users gain a small amount of reputation after an elapsed period of time - T mins. This way newly created identities can start fulfilling contract after T minutes. With an optimal T, this algorithm will create a minimum delay for end-users while effectively deter Sybil attacks. To select the optimal T will require sizable empirical trial however we believe an effective T could be sufficiently small around 10 minutes.

Opportunities

To adequately assess the scale of the opportunities for Tethys network as a general purpose semantic web information network is almost impossible since there has never been such a platform or similar means to access semantic web information. This exercise is somewhat similar to trying to identify the potential market opportunity for search engines before they were introduced. Both technologies primarily focus on improving information dissemination in an information-centric network. Applications for Tethys technology is almost unlimited and most of the areas we can envision today perhaps are just the tip of a iceberg. However, to bring a sense of scale to the readers of this paper, we decided to pick a few different verticals where consumers routinely perform manual searches of deep semantic web information using standard search engines or small-scale aggregators where Tethys network could offer brand new solutions and opportunities.

E-Commerce

Global retail e-commerce sale is projected to reach 4.48 trillion US dollars in 2021.[17] About 65% of online shoppers compare prices online before making a purchase.[18] As we discussed and demonstrated in the Problems section this price comparison process conducted through today's standard search engines is painfully slow and tedious. This perhaps explains why more than 55% of US online shoppers now start their search on Amazon instead of Google. [19] Tethys can offer access to real-time verified semantic product information. It is conceivable a specialized semantic search engine with access to all e-commerce sites, and their products can be built to offer consumer instant insight on price, shipping, ratings, and all other relevant information. In fact, our launching partner Yroo is a company with precisely this mission in mind. Yroo is a big data startup specialized in shopping vertical currently based in Ireland with 11 million USD raised to date.[20] Today Yroo already indexes over 100-million products daily using direct data feed and API integration offered by e-retailers and marketplaces; however they are planning to upgrade a significant part of their pipeline

[17] Global retail e-commerce sales 2014-2021.
https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/
[18] PEW Research Center Online Shopping & Ecommerce Survey.
http://www.pewinternet.org/2016/12/19/online-shopping-and-e-commerce/
[19] BloomReach.
https://www.bloomberg.com/news/articles/2016-09-27/more-than-50-of-shoppers-turn-first-to-amazon-in-product-search
[20] Yroo, The World is Your Store. http://www.nibletz.com/events/collision/yroo

with Tethys. In fact, many of the architectural choices and design innovations came as a result of Yroo's internal prototype - decentralized semantic crawling PoC project. As a launch partner, Yroo has committed a significant amount of resources and technical expertise in partnership with Tethys Network to launch Phase I of the implementation through their browser extension and mobile app environment. The critical objective for Yroo in this partnership is the ability to automatically verify information through Tethys contracts, as well as the ability to control information retrieval frequency hence making sure their information is always accurate and up-to-date. Finally having Tethys information blockchain to record and store historical prices in an immutable way allows Yroo to offer historical prices with confidence and transparency to its consumers. Yroo believes that Tethys tokens could be an powerful incentivization scheme for its users thus helping fueling Yroo's growth in user acquisition and engagement. As a result, Yroo is also planning to launch world's first online store where Tethys tokens are accepted in exchange for digital goods.

Auto Industry

In 2017 US auto industry sold over 17 million vehicles.[21] On average consumers spent 14 hours through the purchase journey with more than 50% of this time spent on researching and shopping online.[22] This is completely understandable for anyone who has purchased a car recently. First of all car buying is a major purchase for most people and the potential saving could easily amount to hundreds or even thousands of dollars. To make matter worse the amount of information and variables to consider for a car nowadays is incredibly large. The followings are some of the major categories of variables the consumers need to consider maker, model, trim level, packages, warranty, accessories, financing, loyalty program just to name a few. Finally to make this research project even more challenging is that auto dealers and automakers run different promotions and special offers in different geographic regions for different periods. The main reason why this research process is so time-consuming is due to the fact that all information above buried in the deep web. The consumers need to manually go to different automaker's website to find out the exact quote for the desired configuration while knowing that the quote might change in a matter of weeks or even days. Facing this kind of time-consuming challenge a specialized search engine or comparison engine, leveraging Tethys Network's unique capabilities, can aggregate all semantic information from the deep web, e.g. from different automakers, dealers, and aggregators website simplifying this research process significantly. Additionally this auto comparison engine will also be capable of recording all of these variable in a chronological order using Tethys blockchain providing an extra dimension of information to consumers that is not currently available; finally, answering the proverbial question, once and for all, "When is the best time to buy a car?" with statistics and hard data.

[21]Vehicle sales in the United States 1977-2017,
https://www.statista.com/statistics/199983/us-vehicle-sales-since-1951/
[22] AutoTrader 2016 Car Buyer Journey Report.
https://b2b.autotrader.com//agame/pdf/2016-car-buyer-journey.pdf

Roadmap

Q3 2017 - Project Tethys Inception

Q4 2017 - Internal Yroo Prototype and Proof-of-Concept

Q2 2018 - Tethys token pre-sale

Q3 2018 - Tethys tokens launch

Q4 2018 - Phase I Tethys Network and VM launch (Standalone & Yroo branded)

Q2 2019 - Phase II launch of Tethys blockchain