Layer 2 scaling solutions

Introduction

Bitcoin and Ethereum implemented a sophisticated mechanism to ensure their designed property: decentralisation and security. Both BTC and ETH1.0 use a power consuming PoW consensus protocol to guarantee the trust less feature. Thanks to the decentralised feature of ETH and its support to a generic EVN, exponentially increasing group of participants start to hold and use ETH. From 2017, ETH became a popular platform for decentralised applications (DApp) supported by smart contract. Because a smart contract requires a significant amount of computational resource, the original design of PoW process inflicts high costs to users. With the development of DApp ecosystem, various types of applications have been deployed on the Ethereum blockchain. At the same time, the high cost became to constraint the development of DApp as long as the value of ETH grew rapidly. In addition, the enormous amount of traffic is overloading the capacity of Ethereum block chain, which aggravated delay and cost [1]. Developers then started to seek for scaling solutions to increase transaction speed and throughput with maintaining the decentralisation property and security.

In this report, ETH is focused as the platform of scaling target. Conceptually there two types of scaling on-chain (Layer 1) solution and off-chain (Layer 2) solution. Layer 1 solution is achieved by upgrading the mechanism of Ethereum protocol. Technology like sharding is introduced in ETH2.0 to reduce the computational burden by splitting the verification work and even code execution to subsets of nodes instead of the whole network [2]. To reduce the computation cost of each transaction, ETH2.0 also applied a new PoS mechanism for verifying transactions [3]. However, it is still not enough to handle the whole traffic, Ethereum community then shifted their focus to Layer 2 scaling solutions.

Layer 2 is a blockchain built on top of Ethereum mainnet that scale up the capacity of Ethereum and shares the security of it. Because the core values of a block chain: decentralisation, security and scalability cannot co-exist perfectly. Layer 2 chains must sacrifice decentralisation or security to a certain extent to accomplish its mission [4]. After a five-year development, various solutions have been put forward and implemented. This report will firstly introduce the historical development of several mainstream Layer 2 scaling solutions, and then give in depth explanations to optimistic rollup and zk-rollup, followed by critical comparisons between these two alternatives and finally a conclusion is made on the prospect of Layer 2 development.

Historical Survey

This report focuses on the attempts of layer 2 scaling solutions on Ethereum network. Technology features and trade-offs of alternative solutions with different concepts and intentions will be introduced. Afterwards, two highly anticipated solutions will be explained in depth and compared.

Layer 2 solutions made different trade-offs on security, transaction cost and efficiency and support to general computing. Mainstream technologies are roll up, side chain, plasma and state channel.

Plasma, inspired by lightning network, was firstly introduced as a Layer 2 scaling option. It reduces the computational cost by creating child plasma chains that are enforced to commit transactions to root blockchain (Ethereum chain) [5]. Because a Plasma chains are only required to keep a record of its local data, the expected computational cost are significantly lower than the original cost. To synchronise the transaction history, a bridge is built to link the Plasma chains and the root chain for submitting commitments. Plasma gained popularity rapidly after the publication of the paper, and it went live in the late 2017. However, as the market gradually calms down, the problems that Plasma introduced began to be brought to public's attention. A lot of effort should be put to detect malicious behaviours of operators. This is a serious departure from the original intent of Layer 2 concept. Long challenge period and mass exit problem of Plasma also contributed to the collapse of Plasma in 2018. Even if plasma cash was invented in mid-2018 to deal with these problems, it also came along with several issues. To ensure the integrity of the transaction, users are required to maintain online in the challenge period and record the entire transaction history of a coin [6]. Lower-than-expected performance and unexpected problems of Plasma led it to a silent in late 2018.

The sidechain approach is similar to Plasma, but the difference is that they are more complete in functionality and have more autonomy. A sidechain is two-way pegged to the root chain that users can send tokens to another chain by depositing on one chain. Most sidechains support full function EVM which behaves exactly the same as Ethereum. Less resource demanded consensus algorithm applied by sidechains made higher throughput and lower cost possible. Although decentralisation is sacrificed to some extent, it is still gaining popularity for its convenience. Lots of products using different implementation of side chain emerged in the past few years, such as Polygon with proof of stake, Skale with Asynchronous Binary Byzantine Agreement and xDai with proof of Authority. Most sidechains use a leader election mechanism for validators to create and verify blocks. This leads to a lower cost than that of a root chain while introducing insecurity and risk of collusion. The drawbacks can be neglected by the users who have strong demand on low-cost Ethereum-like experience. Polygon, previously known as Matic network, was deployed in 2020 and attracted attention of the DeFi area. Most DeFi applications follow in it's development with polygon-compatible versions in 2021. Until now, polygon is still a choice for Layer 2 platform and it's still absorbing new technologies like roll-up to upgrade it's capability. Interoperability applications as a more general form of side chain is also gaining popularity. However, due to the technical features of verification mechanisms used, the side chains attract different groups of investors and DApps. This results in an unorganized ecosystem. In May 2020, polkadot that use a relay chain to connect various parachains was launched to facilitate cross-chain communication. In that year, more developer teams began to build a cross-chain protocols that integrate different technology to meet various needs. In 2021, Connext launched a NXTP protocol for generalized transaction between sidechains [7]. Although side chain provides satisfying scalability, they are still not considered ideal solution because of lack of decentralization and security. As an alternative, rollup is usually considered a promising layer 2 scaling solution.

Roll up is the most anticipated alternative and is going to take up majority of layer 2 market share. There are two major types of roll up: ZK Rollup and Optimistic Rollup. Although they both use the concept rollup by keeping only necessary security information on Layer 1 chain and transfer the computations to Layer 2. The assumptions and security model are substantially different from a technical point of view [8]. The concept roll up was developed by Barry Whitehat in late 2018 which can be found on the GitHub page [9]. He suggested a Plasma-like architecture by transferring activities off-chain and build a communication channel between L1 and L2. Unlike Plasma, transfer of transaction between layers is achieved by algorithm instead of an operator. A validation data called SNARK (Zero-Knowledge Succinct Non-Interactive Argument of Knowledge) is maintained and uploaded regularly to the root chain. Therefore, no extra trust is required because any participants can offer to join the validation process. Shortly afterwards, in September 2018, a refined version called zk-roll-up was proposed by Vitalik. It is stated that zk-roll-up can potentially leverage the processing speed to around 500 transactions per second [10]. This approach provided data availability to public by posting transactions happened on sidechain to root chain along with the roll-up data. In early 2019, projects of zk-rollup were actively developed like zkSync and LoopRing. As an alternative, starkEx developed STARK mechanism on top of SNARK for a better exit guarantee. While the development of zk-rollup was in full swing, another version of roll-up called optimistic roll-up has been tested since 2019. The iconic companies that apply optimistic rollup are Arbitrum and Optimism which has been leading L2 projects in 2022 with market share of 56.45% and 8.72%.

In-Depth Explanation

In this report, these two versions of rollup will be explained in detail, with respect to their design and implementation. Also, the technical feature and main usage scenarios will be introduced followed by several sample projects for demonstration. Zk-rollup will be introduced firstly in the next paragraph.

ZK-Rollup

The word rollup describes the process of packing up a bundle of transaction data and send them all together up to the root chain. And the prefix, zk, stands for "Zero Knowledge" represents the feature of validating message sent by the Layer 2 chain.

In the architecture of zk-rollup, there is a smart contract running in the root chain that keeps the state of transactions happen in layer 2 [11]. And the record is only updated via zero-knowledge validity proof called snark. As for computation, like Plasma, transferred off-chain to reduce cost. On a rollup chain, a batch of transactions are stored in a "batch", which corresponds to block. The job of forging batches and generating snarks are delegated to operators, which are the maintainer of the rollup chain. Also, verifications of snarks are also performed by operators while forging new batches. As the security of rollup chain is not secured as the Ethereum main chain, each transaction happened must be verifiable independent of the trust to operators. Therefore, zero-knowledge proof is the key to implementation of zk-rollup.

Zero-knowledge proof

Zero-knowledge proof was first suggested by Rackoff, Goldwasser and Micali [12]. This primitive version of zk proof method is called interactive proof because the verifier (audience) of the proof is required to ask questions and get feedback from the prover. Afterwards, Blum and his research team designed a framework for non-interactive zk proof (NIZK) [13]. With NIZK, only one communication is required instead of ask and answer iteratively. Later on, protocols such as Pinocchio and Groth16 reduced the verification time to sublinear grade and constrain the argument size to constant [14]. Therefore, it became possible to prove a statement by non-interactive succinct argument with zero knowledge, and that is where the name zk-SNARK comes from. A prover can prove the validity of a statement without revealing any additional information under a Zero-knowledge proof protocol [15]. For example, one can prove that he holds the private of a public key without revealing the value of the key. In zk-rollup, SNARK is used to prove that the account information in a batch is not tempered and update the information stored on the root chain correctly. The whole process of SNARK proof can be abstracted as a circuit, with defined input and public output. In the next part, this report will describe how the information are maintained and modified.

Layer2 components

According to Barry Whitehat [16], In the rollup chain, an array of accounts is stored as leaf nodes in a Merkle tree called account tree. And an account is represented by the hash value of public keys, balance, token type and a nonce. The account tree is initiated as a tree with all leaf node set to none. On each deposit, an account is created, and added to a deposit list. Several unrecorded deposits can be aggregated into a deposit tree which is also a Merkel tree. Then, coordinators add the deposit tree to the account tree by replacing the empty nodes in there. The validity of deposits is then tested by the associated Merkel proof.

To withdraw, an account can initiate a transaction to address 0 which is the reserved address for withdrawal. Afterwards, the smart contract on chain will send the amount received to corresponding account when a SNARK proof is sent by a coordinator.

Once an account is activated through depositing, it can initiate a transfer by specifying the flowing details:

1. from: index of the sender.
2. nonce: Used to avoid duplicated transactions.
3. to: Index of the destination account.
4. amount: Amount to transfer for this transaction.
5. fee: Service fee for coordinators this time.
6. sig: Information related to signature like signature, public point, scalar for the signature, .

Then, a coordinator is required to perform the following operations required by SNARK proof:

1. Verify that the sender address: from with specified public key exists in the account tree by Merkel proof.
2. Match the signature attached with the public key stored.
3. Adjust the balance and nonce.
4. Update account tree by the new account information.
5. Repeat the steps above for receiver: to and obtain the resulting Merkel tree.

This architecture also ensures the data availability by setting all transfer information as public input to the SNARK proof.

Optimistic rollup

According to Offchain Labs, Optimistic rollup use another mechanism for proving the correctness of the current state on layer 2 [17]. Instead of validity proofs of zk-rollup, Optimistic rollups use a fraud proof to verify integrity of off-chain state. The name Optimistic comes from its assumption that a state update is regarded valid by default. That means no associated proof information is required for a node to submit an update. The correctness is guaranteed by other participants in the network who can challenge the updates. Existing agreements are resolved by a disputing protocol that always let the correct nodes win. Also, a reward and punishment system is implemented to incentivise challengers and penalise cheating nodes. Thanks to the reward system, participants tend to behave rightfully. That leads to a lower challenge rate, which keeps the cost of running as low as running the functionalities. The dispute resolution system also leaves a time window for any participants to challenge the submissions. The intention of challenges is basically proving that the commitment is wrong, in another word, is a fraud. If a commitment is not challenged after the time window, it is considered valid [18]. Because complex cryptographic algorithms are not involved fraud proof, there are less constraints on the transaction type and data store. Hence, Optimistic is typically considered a more general solution to layer 2 scaling.

Dispute resolution process

Both Arbitrum and Optimism designate the job of rolling up a batch of transactions to a sequencer (in Optimism) or an aggregator (in Arbitrum) [19]. And each transaction is characterised by a message called Disputable Assertion (DA). A DA describes the final state after a transaction with given initial state and some other context conditions. Along with a DA message, the sender must also stake tokens on the assertion they made. The stake made is then the punishment on cheating and the reward for validators. Once the assertion is proven or disproven, the pledged currency will be transferred to the party who deserve it. Usually, if a fraud is noticed a portion of the stake is burned and the remaining part is paid to validators. After the deadline for challenging an assertion, the DA is settled in the chain and no further challenges are accepted.

Tool chain developed for Optimistic rollup

The main purpose for Optimistic rollup is to build a Layer 2 platform that is compatible to the general applications running on the Layer 1 chain. Because it is not possible to create an identical copy of the existing tools, developers try to make minor modifications them. Therefore, developers can use familiar tools with same workflow and security assumptions. As a result, the optimistic version of following fundamental infrastructures have been implemented: EVM, Solidity, Geth and operating systems. These applications are not only used for executing transactions locally, but also fundamental tools for generating fraud proof.

Fraud proof system

Fraud proof system requires challengers to run the controversial transactions on Ethereum block chain. This requires the transaction message sent by a user to be runnable on Layer 1. However, it is hard to replicate the running context like block information and timestamp from L2 to L1. In order to generate execution results for the transactions, developers for optimistic frameworks provided generic virtual machines that are compatible with Ethereum chain for clients. Nodes can use the virtual machines locally to generate DA messages and challenges and run them on-chain for fraud proof [20]. With the virtual machine, layer 2 applications can move the computations off-chain to save cost significantly.

Different to the NIZK protocol used by zk-rollup, optimistic rollups adopt interactive proving system for fraud proof. The steps of creating a fraud proof using a virtual machine is as follows:

1. Declare the disputing state that is going to be challenged.
2. Synchronise the pre-condition of the doubtful transaction.
   a) Deploy the associated contract on the L1 chain.
   b) Upload the environmental information to L1.
   c) Fetch the data back to L2 virtual machine.
3. Run the transaction.
4. Update the state tree by the new post-execution state.
5. Complete the process and settle with the system.

Because the verification process is ran on L1 and supported by the L2 virtual machine, the integrity of fraud proof is guaranteed by the security of Ethereum chain.

Critical comparison

After a through explanation of the mechanism of both zk-rollup and optimistic rollup, it can be concluded definitely that they have distinct features and trade-offs. Although the market share of optimistic rollup is substantially higher, it still cannot be concluded that OR is a superior choice in every instance. This report will compare these two solutions in different aspects including: generality, running cost and user experience.

Generality

Because of the design of the protocol, optimistic rollups typically have a comprehensive support to Ethereum ecosystem. With the help of L2 virtual machines and the modified tool kits, most existing projects can be migrated to Layer 2 with barely no changes. And because most fundamental functionalities are ready-to-use for developers, three is no extra learning costs to use optimistic rollups. Users can also deploy same smart contracts in an optimistic network without changing the code. Therefore, an optimistic network can simply be regarded as a mimic Ethereum chain running on layer2. However, support to smart contracts is a still a short board for zk-rollup. Because validity proofs are generated by cryptographic algorithms, which is tricky for applications with complex state changes and calculations [21]. Most contemporary zk-rollup applications are function specific platforms specialised in processing transactions. Although the support to general application of zk-rollup is worse than optimistic method, It is certain that zk-rollup is more efficient in some areas such as decentralised exchange. Since late 2021, top projects that apply zk-rollup had started to upgrade to a EVM-compatible version. ZKSync has published ZKEVM for processing smart contracts. In addition, some platform is also seeking a way to integrate different L2 solutions together for to take advantages of the aggregation effect. Polygon has merged Hermez into its portfolio in later 2021. And this team is also developing ZKEVMs in order to be compatible with full opcodes [22]. It can be concluded that optimistic rollup is currently a better option for in terms of generality. But it is possible for zk-rollup to get back to the game after the ZKEVMs are fully launched.

Running cost

The running cost is determined by the design of these two protocols: Optimistic rollup uses “fraud proof” while zk-rollup uses a “validity proof”. As discussed before, validators only need to run challenges to the doubtful transactions. And before submitting the challenge on-chain, they can pre-run the codes locally to avoid high gas fee incurred on L1. The cost of running code locally can be as low as running a normal program. Under a well-designed award & punishment system, it can be assumed that almost all participants behave honestly. Therefore, the overall cost can be considerably low comparing to L2. For the validity proof used by zk-rollup, each batch should be verified by a SNARK. It can be argued that zk-rollup brings a significant computational overhead to each transaction. As a result, validators with intense computational resource will be responsible for the generating SNARKs. In a way, this can be regarded as a threat to decentralisation and transparency [10]. Researchers are still seeking for solutions to reduce computational cost by compressing batches and optimising algorithms. Although in general, the running cost for optimistic rollup is relatively lower, zk-rollup is still an economic choice because the cost is still acceptable comparing to gas fee on L1. Also, some investors concern that the assumptions introduced by fraud proof is not as secure as numerical computations. So, the choice between these two alternatives is still mainly determined by their purpose of using them. If a user is more sensitive to the cost, optimistic rollup is often a better choice.

User experience

User experience is determined by three factors: enter & exit mechanism, ecosystem and learning cost. Both solutions have bridges for linking L1 and L2, users can deposit their tokens on L1 to get L2 tokens, and they can also exit by withdrawing the deposited tokens. In addition to transaction fee incurred, they should also wait for a period until their accounts are confirmed valid. Due to the presence of challenge window in optimistic rollup, investors must wait for typically one week to get funds back [23]. To solve this problem, fast withdrawal services offer a fast exit option. However, users need to pay for extra fees for the service. Even worse, there are various platforms providing a withdrawal service, investors will be bothered by additional learning cost and risk exposure. On the other hand, withdrawing fund from a zk-rollup chain only costs few minutes to be confirmed. Also, withdrawing from a zk chain doesn't need collaterals for backup, this is also beneficial to the liquidity. In terms of ecosystem, and learning cost, optimistic rollup is a better option for normal users. Because most optimistic chains have better compatibilities to Ethereum tool-kits. Users can enjoy a Ethereum-like experience without learning new applications. Zk-rollup, on the other hand, offer a more dedicated service. Users are forced to do more researches and switch platforms for different use case. Because of this, sophisticated users may feel that zk-rollup is handier for professionals.

Conclusion

From the researches on historical Layer 2 scaling solutions to Ethereum blockchain, we can conclude that this area is still in early stage of development where new methods are still being invented. There is also a remarkable characteristic in the development route of L2 technologies is that even divergence in technology lines exist, a higher-level application is built on top of the optimal option. This is a sign that the research is still progressing consistently. Based on the lessons learned from the problems and risks exposed by plasma and sidechains, the rollup approach seems to be a capable to offer a satisfying solution without too much compromise. Currently, both optimistic rollup and zk-rollup are promising alternatives. Although optimistic rollup prevails in market share thanks to a better ecosystem and generality, zk-rollup is still worth waiting for after deployments of mature ZKVMs and toolchain. One absolute advantage of zk-rollup is that it secures transfer by mathematical way instead of by pure simulation-which doesn't sound smart. And less assumptions are introduced by zk-rollup. If the problem of running cost and generality can be solved, zk-rollup will attract more users by its more cryptographic way to rollup. Overall, as the most anticipated L2 scaling solution, rollup has the great potential to solve the scaling problems faced by Ethereum blockchain.

Reference

[1] C. D. I. E. A. E. G. A. J. A. K. A. M. P. S. E. S. E. G. S. D. S. &. R. W. Kyle Croman, “On Scaling Decentralized Blockchains,” Springer, Berlin, Heidelberg, 2016.

[2] E. d. team, “Shard chains,” 3 April 2022. [Online]. Available: https://ethereum.org/en/upgrades/shard-chains/. [Accessed 5 April 2022].

[3] C. Kim, “ETHEREUM 2.0: HOW IT WORKS AND WHY IT MATTERS,” coindesk, 2020.

[4] L. media, “WHAT IS THE BLOCKCHAIN TRILEMMA?,” 15 November 2021. [Online]. Available: https://www.ledger.com/academy/what-is-the-blockchain-trilemma.

[5] V. B. Joseph Poon, “Plasma: Scalable Autonomous Smart Contracts,” 11 August 2017. [Online]. Available: https://plasma.io/plasma.pdf.

[6] H. Q. Ashwin Ramachandran, “The Life and Death of Plasma,” 28 Jan 2020. [Online]. Available: https://medium.com/dragonfly-research/the-life-and-death-of-plasma-b72c6a59c5ad.

[7] B. Arjun, “nxtp: A simpler xchain protocol,” 6 August 2021. [Online]. Available: https://blog.connext.network/nxtp-a-simpler-xchain-protocol-88760697ea04.

[8] Paul, “chaindebrief,” 29 December 2021. [Online]. Available: https://chaindebrief.com/optimistic-zero-knowledge-zk-rollups/.

[9] B. Whitehat, “Roll up,” 5 September 2018. [Online]. Available: https://github.com/barryWhiteHat/roll_up.

[10] V. Buterin, “On-chain scaling to potentially ~500 tx/sec through mass tx validation,” Ethereum developer team, 22 September 2018. [Online]. Available: https://ethresear.ch/t/on-chain-scaling-to-potentially-500-tx-sec-through-mass-tx-validation/3477.

[11] S. Brown, “medium,” 7 July 2021. [Online]. Available: https://medium.com/fcats-blockchain-incubator/how-zk-rollups-work-8ac4d7155b0e.

[12] S. Goldwass, S. Micali and C. Rackoff, “The Knowledge Complexity of Interactive Proof-Systems,” in *STOC '85: Proceedings of the seventeenth annual ACM symposium on Theory of computing*, New York, NY, United States, 1985.

[13] M. Blum, A. D. Santis, S. Micali and G. Persiano, “NONINTERACTIVE ZERO-KNOWLEDGE,” MIT Lab for Computer Science, MA, 1990.

[14] iden3, “Background in ZK,” October 2021. [Online]. Available: https://docs.circom.io/background/background/.

[15] J. Baylina, “circom and snarkJS,” iden3, 2019.

[16] B. Whitehat, “roll_up token: SNARK-based multi-ERC20 side chain,” 11 July 2019. [Online]. Available: https://github.com/barryWhiteHat/roll_up_token#account-leaf-format.

[17] Offchain Labs, “Optimistic Rollups: the present and future of Ethereum scaling,” 18 December 2021. [Online]. Available: https://medium.com/offchainlabs/optimistic-rollups-the-present-and-future-of-ethereum-scaling-60fb9067ae87.

[18] G. Konstantopoulos, “How does Optimism's Rollup really work?,” 29 January 2021. [Online]. Available: https://research.paradigm.xyz/optimism#the-importance-of-software-reuse-in-optimism.

[19] Arbitrum, “Arbitrum Rollup Basics,” [Online]. Available: https://developer.offchainlabs.com/docs/rollup_basics.

[20] H. Kalodner, S. Goldfeder, X. Chen, S. M. Weinberg and E. W. Felten, “Arbitrum: Scalable, private smart contracts,” in *27th USENIX Security Symposium*, Baltimore, MD, 2018.

[21] T. Schaffner, “Scaling Public Blockchains - A comprehensive analysis of optimistic and zero-knowledge rollups,” Center for Innovative Finance, University of Basel, Basel, 2021.

[22] D. Bogdanov, “Optimistic Rollups vs ZK Rollups: Examining Six of the Most Exciting Layer 2 Scaling Projects for Ethereum,” 24 August 2021. [Online]. Available: https://limechain.tech/blog/optimistic-rollups-vs-zk-rollups/.

[23] A. Gluchowski, “Optimistic vs. ZK Rollup: Deep Dive,” Matter Labs, 4 November 2019. [Online]. Available: https://blog.matter-labs.io/optimistic-vs-zk-rollup-deep-dive-ea141e71e075#:~:text=Due%20to%20the%20problems%20mentioned,week%20fraud%20proof%20challenge%20window..