Generaitiv Whitepaper

@undervaluedjpeg @Ada0x05 @zk_brain

April 17, 2023

1 Introduction

Today's AI Models are used to generate an increasingly diverse set of media and outputs. From images and text to video and audio, the possibilities of what can be generated is ever increasing. Also increasing is the number of users directly interacting with centralized AI gatekeepers. Such entities as OpenAI, Midjourney and Stability AI have amassed millions of users in a short number of months, with billions of inferences being run in centralized data centers.

Growth in this area is expected to continue as models increase in their size and complexity, allowing for continually better output and use cases. However with such an amassing of a new market, one must closely examine the incentives and business models of the centralized entities and seriously ponder alternatives. As seen in the web2 space, open source always beats closed source software and the same will apply to AI models.

2 State of the Industry

With the successful deployment of the Ethereum Proof of Stake (PoS) update, many GPU operators have been left with excess GPU computing power. Many operators have switched to mining other crypto currencies, but receive a lower rate of return compared to Ethereum Proof of Work (PoW) mining. GPU operators are actively seeking new opportunities for revenue generation to make use of their upfront hardware investments.

Generative AI technology has shown its potential in both consumer and industry contexts, with a growing number of users directly interacting with AI models. The need for on-demand AI inferences will continue to grow as user adoption grows.

3 Generaitiv's Proposal

Generaitiv is proposing a decentralized GPU network run by its community to complete AI based workloads. Node operators are incentivized to participate and secure the network. Users of the network will request work to be completed. The network will produce useful work with the GAI token as a representation of GPU compute time. Operations for this network will occur on a low cost L2 network with a bridge to move tokens between network layers.

4 Token Staking

Staking is a commonly used mechanism to both reward holders and provide support to a platform ecosystem. Holders will lock their tokens for set intervals of time (a reward period) and the pooled rewards will be claimable to holders in proportion to their staked tokens. In practice this becomes the following equation.

$$S \sum_{i=k}^{n-1} \frac{R_i}{T_i}$$

Where:

S is the constant amount staked over the total reward period by a holder

n is the reward period the holder withdraws their stake

k is the reward period the holder starts their stake

i is the reward period

R_i is the variable reward of the reward period i (the protocol fees during staking period i)

T_i is the total amount staked at a given reward period i

4.1 Staking Rewards & Incentives

As part of the 30% locked tokens, 5% has been allocated for an additional staking incentive to be distributed over a one year period.

As of April 2023, the generaitiv protocol generates revenue from NFT minting fees and secondary marketplace fees. 50% of these collected fees will be allocated as rewards for stakers. Please refer to the attached figure at the end of this document for a full picture of the flow of funds.

4.2 Fiat Gateway

It is acknowledged not all parties who wish to utilize the L2 GPU network will also be interested in acquiring GAI on L1, bridging to L2 and then performing their transactions. Generaitiv will facilitate a direct fiat to GAI L2 token gateway via an L2 liquidity pool backed by staked L1 tokens. Collected fees associated with this gateway will contribute to the rewards for stakers.

Generaitiv requires two tiers of staking to ensure all cohorts of holders can participate in securing the network. Tier 1 represents the segment of users who hold $GAI and either do not own GPU capable hardware or do not want the burden of running and maintaining such hardware. Tier 2 represents the segment of users who hold $GAI and will contribute useful GPU based work to the network.

4.3 Tier 2 - GPU nodes & Distributed Network

Tier 2 nodes will register on the network by staking a minimum threshold of $GAI. Nodes will receive rewards proportional to the amount of useful work successfully completed and delivered to the network. Tier 2 nodes will be responsible for operating and maintaining their hardware, updating to the latest software clients, and ensuring adequate network connectivity to receive and deliver workloads.

Tier 1 stakers can delegate their tokens to a trusted Tier 2 node, resulting in a higher expected reward.

5 AI Workloads

There are many types of AI workloads which the network will complete. Open source models are transparent to everyone on the network. Since the output of these models is deterministic, any node which has knowledge of the exact input & selected model can perform the same operation and compare the output. Model Provenance is determined by using the hash of the entire repository of files, scripts, and processes required to run a specific model. This will ensure nodes operated by different parties would achieve the same output from the same input.

Due to the nature of running AI models and switching costs associated with operations (bandwidth, time to load to GPU VRAM, storage wear and tear) nodes may opt to run a subset of all models supported by the network. It is also expected some models may be in higher demand than others, leading to market based pricing for timely and prioritized execution of jobs.

Nodes will be incentivized to run the models most demanded by users of the network due to the open market nature of the network. Running higher demand models will result in more available workloads for completion.

5.1 L2 Network Nodes

The network has three types of nodes. Consumers, Producers, and Validators[1]. Details of each node type are found below.

5.1.1 Consumer

This node type requests work to be done. The Consumer posts a bounty and instructions for the job to be completed. The Consumer can select an available Producer for the job or leave it open for any node on the network. This is also useful for batching operations, where a Consumer would like many variations of a job, as it would allow parallel completion of the generations by the network.

Once the Producer has completed the job, the Consumer has the opportunity to review the work. If the work is acceptable, the Consumer releases the bounty to the Producer and the transaction is completed. The Consumer does not need to review the work right away, but after X hours the Producer can claim the bounty as it is assumed to be correct by default. The Consumer receives a small incentive, in the form of a percent discount, when they are quick to approve the work and release the bounty.

5.1.2 Producer

This node type completes jobs and collects bounties upon successful job completion. To become a Producer, one must stake a minimum threshold of tokens. Producers are required to complete work consistently and timely. Producers also receive direct incentive (tips) from Consumers to ensure timely delivery of completed work.

Not all work completed by Producers will be done faithfully. If a Consumer believes a Producer's work to be incorrect, they will ask the network to choose 2n Validators (where n is any natural number) to review the work and submit a vote on what the output should have been. If the work is incorrect, the Producer will have a portion of their staked tokens slashed. Slashed tokens are awarded to the Validators who confirmed the output. A portion of the slashed tokens will also be burned.

To prevent malicious Consumers from forcing an unneeded review of a Producer's work, they must post additional collateral along with the claim. If it is found the Consumer's claim of bad output is incorrect, the Consumer will lose this additional collateral and it will be distributed to the Validators for their efforts.

[1] While Producers and Validators are presented as distinct node roles, it is expected the technical requirements of operating either node type to significantly overlap. (i.e. all Producers can act as Validators and vice versa)

5.1.3 Validators

This node type can be called upon to confirm the execution of Producer nodes. To become a Validator, nodes must stake a minimum threshold of tokens. Using the job output from the Producer as a seed, the network deterministically selects Validators who will rerun the job and use a hash of the output as their vote.

Validators who do not vote with the majority of selected Validators or Validators who fail to respond during the dispute claim window will be considered bad actors and will have a portion of their tokens slashed and returned to the other validators.

5.2 Sample Good Transaction

A Consumer posts a job up with a bounty. A job contains the instructions and steps for the generation to be completed. A Producer node picks up the job, performs the calculation and returns the work via pinned IPFS link. Once confirmed by the Consumer, the Producer is awarded the bounty. Job well done.

Several optimizations could be made to make the release of funds more efficient, such as using zk-proofs on a zkEVM to batch transactions.

5.3 Sample Bad Transaction

A Consumer posts a job up with a bounty. A job contains the instructions and steps for the generation to be completed. A Producer node picks up the job, performs the calculation incorrectly and returns the work via pinned IPFS link. The Consumer notices the output is not as expected and posts a claim against the Producer while also posting additional collateral to the bounty. The additional collateral is proportional to the number of Validators requested to review the work.

The Validators are selected deterministically based on the hash of the output, the time of the block of output submission, and a hashed secret submitted by the Consumer in the original job details. The Validators have a set period of time to respond. Once the set time has elapsed, or all Validators have submitted their votes, a tally of votes is completed and a decision is made. Since all Validators have voted against the Producer, the Producer is slashed and the tokens are shared by the participating Validators.

Because the submitted votes are also the correctly completed work, the Consumer has received the correct output for their original request and their bounty is paid as a bonus to the Validators. The Consumer has their additional collateral returned. Job well done!

5.4 Sample Transaction with non-unanimous Validator response

In the unlikely event the selected Validators are unable to come to agreement on the output for a given job, an additional set of Validators will be selected and the claim process will be repeated. If this additional vote is unable to come to resolution, generaitiv as an authority will have final say and determine the correct output and then deliver it to the Consumer. The Consumers additional locked collateral will be released back to them.

This also acts as a fallback in the case of an underlying system error where despite the deterministic output of a model for some input honest validators end at different results due to software or hardware differences. This is also an unlikely scenario.

6 Party Validator Coordination

Since Validator selection is deterministic, there may be scenarios where either party can make false claims or submit bad information knowing a friendly set of Validators would be selected and they will vote dishonestly in their favor. This would result in an incorrect slashing and reward to one or more parties.

Two such scenarios are possible and mitigated by the protocol. First scenario; a Consumer falsely claims the Producers work was invalid, knowing the Validators will vote in the Consumers favor. Second scenario; a Producer submits incorrect work knowing when the Consumer posts a claim against the job the selected Validators will be friendly towards the Producer.

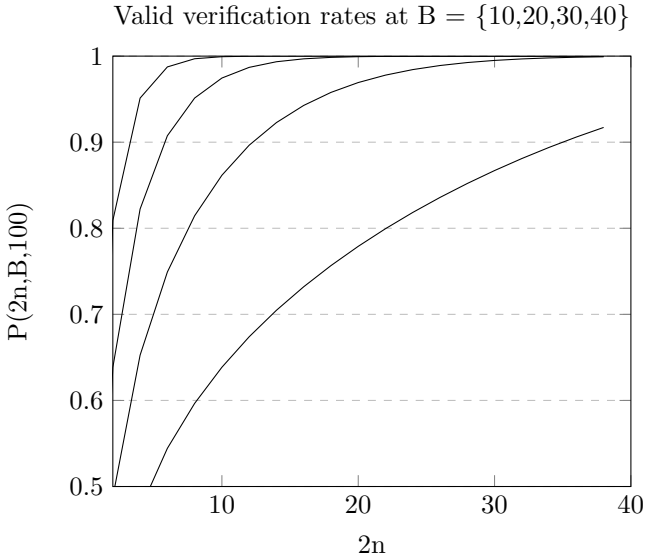
The first scenario is mitigated by the time of submission by the Producer. It is not possible for the Consumer to know beforehand exactly which block the Producer will submit it's completed workload. This is crucial as the Consumer could target an honest Producer by precomputing the correct output hash and selecting a secret to submit with the job which will result in a friendly Validator selection upon claim submission.

The second scenario is mitigated by the hashed secret submitted by the Consumer during job submission. When the Consumer posts a claim, they must also reveal the secret. A malicious Producer is unable to target a Consumer by either submitting a specific hash or timing it's submission to get a favorable Validator selection.

These two mitigation features in tandem act to keep both the Consumers and Producers secure, without needing to trust each other.

Generally more active Validators on the network is better assuming they are operated by arm's length parties. This increases the probability of random and unrelated Validators selection during a claim process.

6.1 Probability of correct validation



Assuming random Validators are selected, the probability a Validation will have a minority of bad actors process it is given by the equation

$$P(2n, B, N) = \sum_{b=0}^{n-1} \binom{2n}{b} \cdot \frac{B!}{(B-b)!} \cdot \frac{(N-B)!}{(N-B-2n+b)!} \cdot \frac{(N-2n)!}{N!} \qquad (1)$$

Where:

2n is the number of Validators chosen for said validation
B is the size of the largest group of coordinated bad actors
N is the total number of Validators in the network

With a 100 Validator network and 20 bad actors it only takes a validation pool of 12 for a roughly 99% success rate. With a network of 422 Validators and 20 coordinated bad actors the chance of an unfavourable Validator selection is $\frac{1}{2^{50000}}$ (A 3x network size increase from here reduces the probability to roughly $\frac{2}{10^9}$).

This is seen as a good tradeoff between certainty of job completion and overall network efficiency. It is unrealistic and impractical for all nodes in the network to calculate the output of all jobs and implement a typical consensus model. Some emerging methods such as Multiparty Compute (MPC) offer efficiency compared to a typical method but create overhead in time and bandwidth. Therefore a new mechanism is required.

Upon bootstrapping a new type of network, it may take some time for adoption and a sufficient number of nodes to be on the network. Generaitiv may choose to

supply sufficient nodes as adoption scales and stabilizes, and modify parameters of the network to support operations.

6.2 Node Reputation

The protocol will also track the reputation of each node. A node's reputation will be used to determine selection in the processing of jobs, in both a Consumer and Producer node type.

7 Conclusion

Generaitiv envisions a world where the best work being done in AI is community driven and open source. The overwhelming advantages of open source AI models, including increased innovation, collaboration, accessibility, and adaptability, make them more likely to outperform their closed source counterparts in the long run. By fostering a global community of diverse contributors, open source AI projects harness the collective intelligence of thousands of minds, leading to a rapid and organic evolution of algorithms and applications. This democratization of AI technology not only accelerates progress but also helps mitigate the risks associated with monopolistic control of AI advancements.

By embracing the open source philosophy, we will ensure a more equitable, efficient, and innovative future in the realm of artificial intelligence, ultimately propelling humanity towards unprecedented technological heights. While it must be said this is a bold vision, in the long term it is abundantly clear this is the only path forward.

Here's to an open source AI future.

8 Footnote on Decentralized Training

The generaitiv team believes there is an important distinction to be made between decentralized and distributed networks currently being built and their implications on training the next generation of AI models. Today's transformer and diffusion models training techniques rely heavily on specialized hardware, often not even available for purchase by a typical hobby level GPU enthusiast.

While many techniques are used to improve model performance and training time, the most common methods today are 1) Larger datasets 2) Optimized methods to reduce training time and 3) Training for more epochs. Reviewing these:

This makes no operational difference when decentralized, assuming the dataset is publicly available and easily obtainable by all parties. If the training nodes are not co-located, as would be expected in a decentralized and distributed network, there may be increased bandwidth costs to access the datasets since each node may have to individually access the dataset. Many optimizations are being made at a very low level. Full utilization of GPU VRAM vs host machine RAM, small timing and cache improvements. These improvements save nano-seconds. All of these may be small 1-2% improvements but add up when considering the number of operations required for training. In general, running training for longer leads to better model performance. In combination with the above improvements, better models can be achieved in the same amount of training time as before when the underlying operations are completed faster.

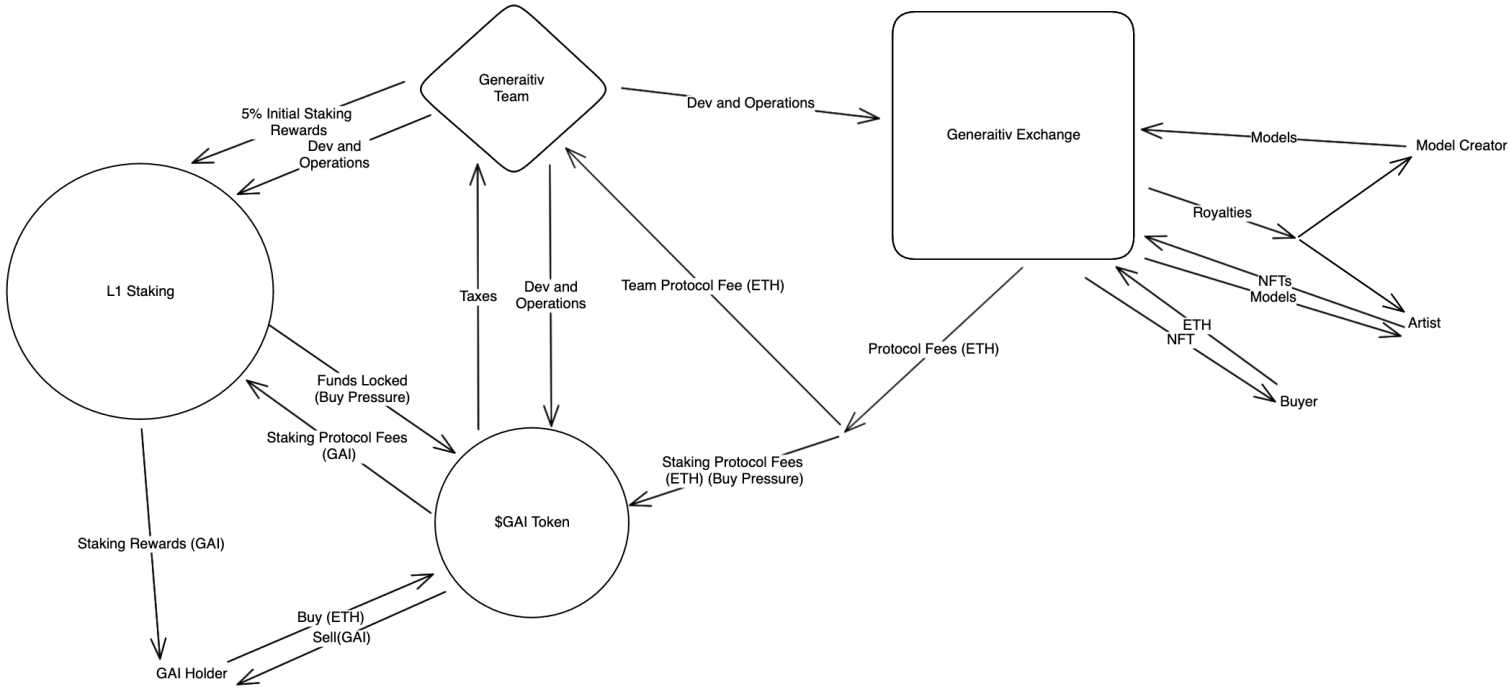
8.1 The 40k problem

The earth has a circumference of 40,000 km. In a fully realized decentralized and distributed network, this becomes a coordination and latency problem. This is not a limit of today's communication technology but an acknowledgement of the laws of physics, more specifically the speed of light. The latency of communication between nodes becomes problematic and will not scale with today's training techniques.

It is crucial to view decentralization in a different way. Decentralization - we do not want all the GPU training power in the hands of a few large corporations. Rather we want many nodes, owned and operated by unrelated parties, to create a market where any party who wants to train a model can.

Generaitiv aims to create protocols and networks which have the incentives designed such that over time this is the resulting structure.

Generaitiv EcoSystem post L1 Staking



Generaitiv L2 Staking EcoSystem