MatrixAI Litepaper

Contract Lab // Released v0.1 // 2023-07

Abstract

The progress of AI heavily relies on large-scale training computation, but the high resource requirements pose barriers to access and impede progress. Current solutions for computational resources are monopolistic, expensive, and impractical for large-scale AI. Decentralized computing services offer advantages such as cost reduction, privacy protection, and innovation opportunities. This paper introduces MatrixAI, a decentralized AI computing marketplace that aggregates idle computing resources worldwide. MatrixAI aims to be the AI computing resource layer network of the Web3 era, supporting diverse scale requirements and breaking centralized monopolies. By empowering computing suppliers and providing cost-effective solutions, MatrixAI shapes the future of decentralized computing and fosters AI innovation.

1 Introduction

1.1 Background

In the past five years, many significant advancements in deep learning have been achieved through the utilization of increasingly large training computation [2,3]. Such large-scale training is accomplished by simultaneously employing hundreds or thousands of specialized accelerators with high internal chip communication bandwidth, such as Google TPUs, NVIDIA A100 and H100 GPUs, or AMD MI250 GPUs. These accelerators are utilized for several weeks or months to compute thousands or millions of gradient updates. Consequently, the tremendous resource requirements involved in constructing these foundational models pose significant barriers to accessing them, and without a means to capture value while sharing resources, it could potentially lead to stagnation in AI progress.

Currently, the solutions available for providing computational resources are either monopolistic and expensive or impractical due to the complex computations required for large-scale AI. Meeting the ever-growing demand necessitates a cost-effective utilization of all available computing resources. The present challenge lies in the limitations imposed by the asymptotic progress of microprocessor performance on the computational resources themselves, compounded by chip shortages resulting from supply chain and geopolitical factors.

On the other hand, the average utilization rate of global cloud computing data centers has consistently remained low, indicating the presence of a substantial amount of idle computational resources. Furthermore, with the continual rise in computing power of consumer-grade devices, there is untapped potential in the form of idle computing resources from personal computers, servers, and mobile devices used by individuals and businesses. Additionally, decentralized data transmission and storage infrastructures like IPFS, Filecoin, and Storj are constantly improving.

1.2 Motivation

The catalyst for each wave of technological innovation is often the transformation of something expensive into something affordable enough to be wasteful.

Currently, in the field of physical infrastructure, there is a dominant market controlled by vertically integrated giants such as AWS, GCP, Azure, Nvidia, Cloudflare, Akamai, among others. These companies enjoy high-profit margins within the industry. This situation results in high computational costs for new entrants in the AI field, particularly in the LLM domain, which hampers the development and widespread adoption of AI technologies. However, decentralized computing services offer numerous advantages, including decentralized resources, elasticity and scalability, cost reduction, privacy protection, high reliability, as well as opportunities for innovation and collaboration. We firmly believe that decentralized computing services will be the key to overcoming the current high cost of computational power, making it affordable and accessible, thus opening the doors to technological innovation in the AI industry and paving the way for the future of the AI era.

In order to achieve this goal, we have taken the first step by establishing a decentralized AI computing marketplace called MatrixAI. Our aim is to aggregate idle AI computing resources from around the world. The vision of MatrixAI is to attract global computing suppliers to participate in the network through a fair and transparent incentive mechanism, thereby creating a vast pool of idle computing resources. We envision MatrixAI as the AI computing resource layer network of the Web3 era, providing support for small-scale AI computing services and high-performance computing clusters to meet diverse scale requirements.

MatrixAI is dedicated to breaking the current centralized monopoly, bringing innovation and progress to AI applications across various industries, and promoting greater openness and sustainability in AI computing services. We firmly believe that through the efforts of MatrixAI, computing suppliers worldwide will be able to unleash their full potential, while computing demanders will gain access to more flexible, efficient, and cost-effective AI computing solutions. We look forward to working with you in shaping the future of this promising decentralized computing field.

2 Goals

As a decentralized AI computing infrastructure, MatrixAI is committed to achieving the following goals.

2.1 More Economical Supply

Providing Equal Opportunities for Hardware Suppliers to Become Service Providers. It establishes a marketplace where anyone can join as a "miner" and exchange their CPU/GPU computing power for economic rewards, thus introducing competition to existing providers. While companies like AWS undoubtedly enjoy a 17-year head start in terms of user interface, operations, and vertical integration, MatrixAI attracts a new user base that cannot accept pricing dictated by centralized suppliers, aiming to serve a price-sensitive demographic.

2.2 Benign Subsidy Mechanism

Creating a Competitive Market to Reduce Customer Payment Costs. In comparison, AWS EC2 requires approximately 55% profit margin and a 31% overall profit margin to sustain its operations. The token incentives/block rewards provided by the MatrixAI network serve as a novel source of revenue. This incentive mechanism is designed to attract more computing power providers to join the network, as they have the opportunity to earn additional income by participating in computational tasks. Such motivation will increase competition among computing power providers, driving them to offer more competitive prices and higher-quality computing power services. Therefore, by introducing more computing power providers and incentive mechanisms, the computing power trading market can achieve supply-demand equilibrium and drive competitive reductions in computing power prices.

2.3 Verifiable Computing Power

The verifiability of computing power within the network is a crucial safeguard for the orderly and transparent nature of the computing power trading market. This characteristic ensures fairness and trustworthiness in computing power transactions. On one hand, honest computing power providers will receive more reasonable economic rewards as their provided computing power can be accurately verified and assessed. This encourages computing power providers to offer high-quality computing power services and establish a good reputation.

Furthermore, in order to ensure that computing power buyers can obtain more valuable metrics as references, the computing power trading market needs to disclose the real computing power situation of each computing power provider. This can be achieved through open verification mechanisms or third-party audits. The transparency of the computing power providers' real computing power situation helps computing power buyers make informed decisions and choose the computing power resources that best suit their needs. Such transparency also helps prevent false advertising or fraudulent activities, thereby maintaining the stability and reliability of the entire computing power trading market.

3 Design

3.1 Roles and Architecture

MatrixAI Network is a decentralized AI computing infrastructure based on Substrate [1]. It encompasses various roles within its ecosystem.

3.1.1 User

Users with Training Model Requirements.

3.1.2 Trainer

Any user with idle computing power resources can join the MatrixAI Network as a trainer without any barriers to entry. Trainers, as consensus nodes within the network, can earn block rewards by contributing valid computing power. Valid computing power can be accumulated through two categories.

- *Drilling*: Accomplishing measurable computational tasks automatically assigned by the network. Trainers can seamlessly process such tasks upon joining the network. These computational tasks are released by projects collaborating with the MatrixAI Network community and typically hold practical application value. The published tasks fall within the realm of machine learning

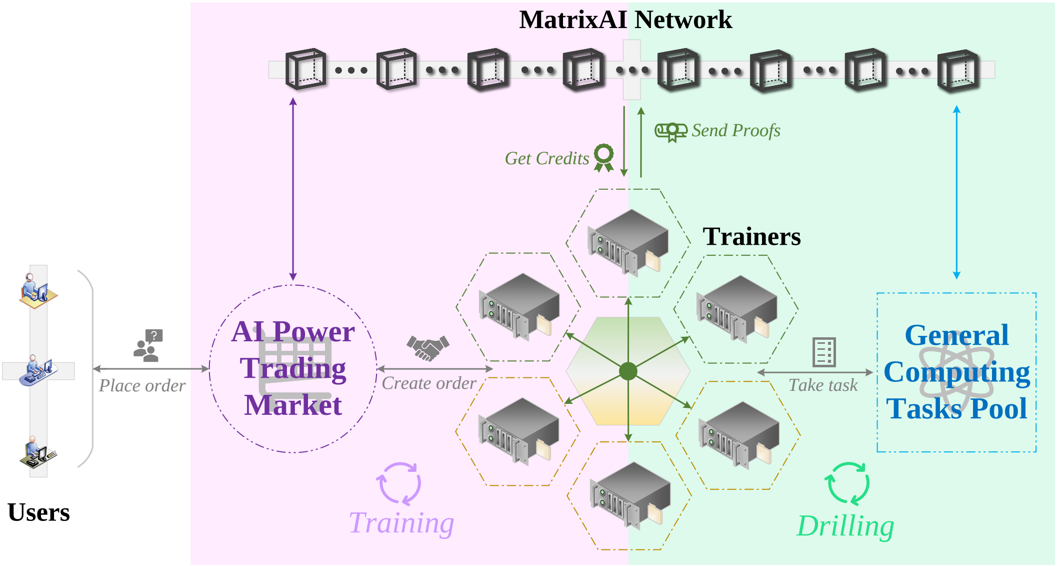

Figure 1: System Architecture of MatrixAI

and can be used to estimate the trainers' actual computing power. Additionally, due to their state independence, these tasks can be easily divided and verified, making them suitable for machines with varying hardware conditions.

- *Training*: By listing computing power resources for sale on the computing power trading market, trainers enter into commitments with users and complete the intended model training. Trainers who join the network have the flexibility to list their computing power resources on the computing power trading market at any time. Before initiating an order and determining the pricing, trainers can refer to the actual conditions of the computing power market. When users select training machines, they can browse through the reported key hardware configuration information of the machines, as well as their historical valid computing power values as references. Once the order is finalized, trainers download the required data from the location specified by the user and proceed with model training as instructed.

3.2 Proof of Hashrate

Proof of Hashrate (PoH) is similar to Bitcoin's PoW (Proof of Work) in the sense that both rely on the workload of consensus nodes to compete for block generation within each cycle. However, there are notable differences. In PoH, trainers do not need to perform a large number of hash computations to find a hash value that satisfies certain conditions. Instead, they can calculate their valid computing power values through Drilling and Training. Additionally, PoH incorporates the VRF (Verifiable Random Function) algorithm [6] to bind the valid computing power value of each trainer with the weight of the VRF. Trainers with higher computing power values have a higher probability of generating a random number that meets the criteria, thereby gaining block generation rights and associated rewards.

3.3 Computing Power Evaluation System

The computing power evaluation system is the cornerstone of the PoH consensus mechanism, ensuring the orderliness and transparency of the computing power trading market.

The computing power evaluation system will dynamically assess the computing power of each trainer, represented numerically as the effective computing power value. The evaluation criteria are based on the quality and quantity of completed Drilling and Training tasks in the trainer's historical records.

To incentivize trainers to contribute more computing power resources to the network and complete a greater number of computational tasks, PoH utilizes the effective computing power value as the VRF weight for each node. Trainers with higher computing power values will have a higher probability of obtaining block generation rights and associated rewards.

Given that the hardware configurations of trainers are not externally perceptible, it is challenging to directly rely on unilaterally provided hardware specifications when trainers set their prices. To provide users with more valuable metrics for reference, the computing power trading market will publicly disclose the effective computing power values of each trainer.

The computing power evaluation system calculates the effective computing power values separately for Drilling and Training tasks, denoted as D_{ECP} and T_{ECP}, respectively. Their relationship can be expressed as: $Total_{ECP} = 0.4 \times D_{ECP} + 0.6 \times T_{ECP}$.

3.3.1 Definition of Computing Power Value

To make computing power resources measurable, we have designed a set of rules for quantifying computing power values. Under these rules, the minimum unit of computing power value is defined as "ut," which represents the amount of CPU time required to execute the Whetstone benchmark test at a rate of 1,000 MFLOPS on a reference computer for 1/200th of a day. For example, if a machine has a computing power value of 200 ut, it indicates that it has performed computational tasks equivalent to a full day of workload on a 1 GigaFLOPS benchmark computer (including both Drilling and Training tasks).

In MatrixAI Network, two computing power values are maintained.

- *Total Computing Power Value*: The sum of computing power values obtained by a trainer since joining the network.
- *Effective Computing Power Value*: The average computing power value obtained per day over a recent period of time. This average value is reduced by half every week.

3.3.2 D_{ECP} Scoring Rules

Drilling tasks are released by collaborative projects within the MatrixAI Network community. Each task is assigned a difficulty level by the collaborating project and is validated by the community. The difficulty level is measured in terms of computational power required, represented by the algorithmic complexity or workload. Therefore, the scoring rules for D_{ECP} are as follows.

$D_{ECP} = The\ Task\ Difficulty\ \times The\ Number\ of\ Tasks$

3.3.3 T_{ECP} Scoring Rules

Training tasks are assigned to trainers based on their acceptance. The scoring rules for T_{ECP} are similar to those of D_{ECP}, as they are determined by the difficulty of the tasks. However, the difference lies in the fact that the difficulty of T_{ECP} tasks is estimated based on the total number of floating-point operations required for model training.

3.4 Proof of Drilling

The design of PoD aims to verify whether trainers have faithfully completed Drilling tasks and grant them the D_{ECP} score.

The basic principle of PoD verification is to achieve consensus by having multiple trainers complete the same task and return the same work units. If they all reach a consensus, the computational power will be calculated, and all trainers will receive the same amount of credit, regardless of their hardware conditions.

On the contrary, if multiple trainers produce different results for the same Drilling task, all participating trainers will lose the corresponding computational power value.

3.5 Proof of Training

The essential requirement for model training outsourcing services is to ensure the authenticity and reliability of the training process.

In a decentralized computing network, users should not blindly trust that the training providers will perform their work faithfully. On the contrary, trainers often violate agreements and commitments in pursuit of profit. For instance, trainers may engage in arbitrary behavior, deviating from the required training process and providing users with erroneous model data.

In the crypto world, we typically adhere to the principle of "don't trust, verify it."

Jia et al. proposed Proof of Learning (PoL) inspired by research on Proof of Work and verifiable computation [4]. PoL utilizes metadata from the gradient-based optimization process to construct a certificate of work completion, providing evidence that the training party has performed the necessary computational work to obtain a set of model parameters correctly.

Zhang et al. identified the vulnerability of PoL to "adversarial samples" and demonstrated, both theoretically and empirically, their ability to generate an effective proof at significantly lower cost than what the prover would require [7]. Building upon PoL, Zhang et al. introduced a training traceability scheme in 2023, leveraging intermediate checkpoints saved during the model training process to create a coherent chain of models as ownership certificates [5].

Shavit proposed a lightweight activity logging strategy based on chip firmware, enabling monitoring of the chip's behavior [8].

Building on the aforementioned approaches, we have conducted further in-depth research and introduced Proof of Training (PoT). The principle behind PoT is to validate whether the intermediate checkpoints generated during the model training process align with the resulting model output through empirical comparisons, such as verification accuracy and parameter distribution distance.

With the support of PoT, the training party only needs to perform model initialization and sequentially save the checkpoints from each training round as a coherent proof bundle during the regular model training process. Anyone who obtains the proof bundle can act as a verifier. Utilizing a series of validation algorithms, which we refer to as the "full-pass rules," we can verify whether the training party has faithfully completed the model training.

The "full-pass rules" consist of the following six verification conditions, and only when all conditions are satisfied, the verification is considered successful.

- *The monotonicity of verification accuracy*: Given a validation dataset D_{val} that is similar to the training dataset, the verification accuracy of each checkpoint on D_{val} should be monotonically non-decreasing.

- *Parameter distribution continuity*: Assuming the model is trained with a sufficiently small learning rate, for any two adjacent checkpoints C_i and C_{i+1} and a small threshold δ, the weights in all layers should satisfy a certain condition.
- *Initial parameter distribution*: The parameters of the initial model should follow the desired Gaussian Mixture Model (GMM) distribution. Given the initial model C_0, the weights in all layers should satisfy a certain condition.
- *Independence of initial parameters*: The parameters of the initial model should be independent. For the weights w in the initialization layer, if we consider two different parameters w_i and w_j as random variables, their covariance should be 0.
- *Monotonicity of weight distance*: Assuming the model training converges properly, the distance between intermediate checkpoints and the converged model should monotonically decrease to zero.
- *Small distance between initial and converged models*: Assuming the deep neural network model is sufficiently complex, the distance between the converged model and the initial model in the same model chain is likely to be much smaller compared to the distance between the converged model and other randomly initialized models.

References

[1] The blockchain framework for a multichain future. *https://substrate.io/*.

[2] Scaling laws for neural language models. 2020.

[3] Jordan Hoffmann et al. Training compute-optimal large language models. 2022.

[4] Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1039–1056. IEEE, 2021.

[5] Yunpeng Liu, Kexin Li, Zhuotao Liu, Bihan Wen, Ke Xu, Weiqiang Wang, Wenbiao Zhao, and Qi Li. Provenance of training without training data: Towards privacy-preserving dnn model ownership verification. In *Proceedings of the ACM Web Conference 2023*, pages 1980–1990, 2023.

[6] Silvio Micali, Michael Rabin, and Salil Vadhan. Verifiable random functions. In *40th annual symposium on foundations of computer science (cat. No. 99CB37039)*, pages 120–130. IEEE, 1999.

[7] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[8] Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*, 2023.