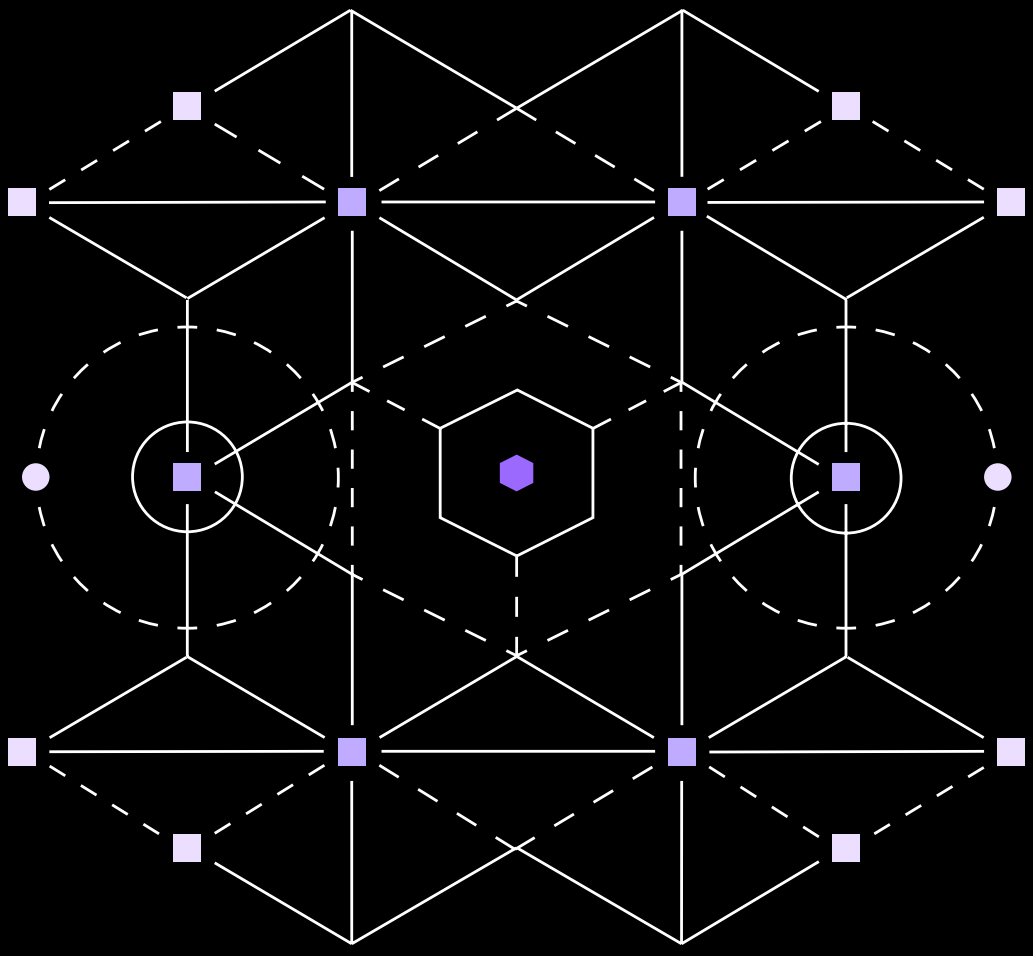
galaxy

Galaxy Research

Understanding the Intersection of Crypto and AI

FEBRUARY 14, 2022



Author & Acknowledgements

Lucas Tcheyan
Galaxy Research

This report is a product of Galaxy Research, a research organization within Galaxy, the leading provider of financial services in the digital assets, cryptocurrency, and blockchain technology sector. Galaxy Research provides top-tier market commentary, thematic views, tactical insights, and deep protocol research.

This report was written February 2024.

View our publicly available research at www.galaxy.com/research. Contact us at research@galaxy.com.

Contents

Introduction

The advent of public blockchains is one of the most profound advancements in computer science history. But the development of artificial intelligence will, and already is, having a profound impact on our world. If blockchain technology provides a new template for transaction settlement, data storage, and system design, artificial intelligence is a revolution in computation, analysis, and content delivery. Innovation in the two industries is unlocking new use cases that could accelerate adoption of both in the coming years. This report explores ongoing integrations of crypto and AI with a focus on novel use cases that attempt to bridge the gap between the two, harnessing the power of both. Specifically, this report examines projects developing decentralized compute protocols, zero-knowledge machine learning (zkML) infrastructure, and AI agents.

Crypto provides AI with a permissionless, trustless, and composable settlement layer. This unlocks use cases such as making hardware more accessible through decentralized compute systems, building AI agents that can execute complex tasks requiring the exchange of value, and developing identity and provenance solutions to combat Sybil attacks and deep fakes. AI brings to crypto many of the same benefits we see in Web 2. This includes enhanced user experience (UX) for both users and developers thanks to large-language models (i.e., specially trained versions of ChatGPT and Copilot) as well as the potential to significantly improve smart contract functionality and automation. Blockchains are the transparent data rich environments that AI needs. But blockchains also have limited computational capacity, a major obstacle to direct integration of AI models.

The driving force behind ongoing experimentation and eventual adoption at the intersection of crypto and AI is the same that drives much of crypto's most promising use cases - access to a permissionless and trustless coordination layer that better facilitates the transfer of value. Given the enormous potential, participants in the space need to understand the fundamental ways in which the two technologies intersect.

Key Takeaways

- Crypto and AI integration in the immediate future (6 months to 1 year) will be dominated by AI applications that enhance developer efficiency, smart contract auditability and security, and user accessibility. These integrations are not specific to crypto but enhance the on-chain developer and user experience.
- Decentralized compute offerings are implementing AI-tailored GPU offerings just as there is a significant shortage in high-performance GPUs, providing a tailwind for adoption.
- User experience and regulation remain obstacles to onboarding decentralized compute customers. Recent developments at OpenAI as well as ongoing regulatory reviews in the United States, however, highlight the value proposition of permissionless, censorship resistant, decentralized AI networks.
- On-chain AI integrations, especially for smart contracts capable of using AI models, require improvements in zkML technology and other computational methods that verify offchain compute on-chain. Lack of comprehensive tooling and developer talent as well as high costs are barriers to adoption.
- AI agents are well suited for crypto where users (or agents themselves) can create wallets for transacting with other services, agents, or people. This is not currently possible using traditional financial rails. Additional integration with non-crypto products is needed for broader adoption.

Terminology

Artificial Intelligence is the use of computation and machines to imitate the reasoning and problem-solving abilities of human beings.

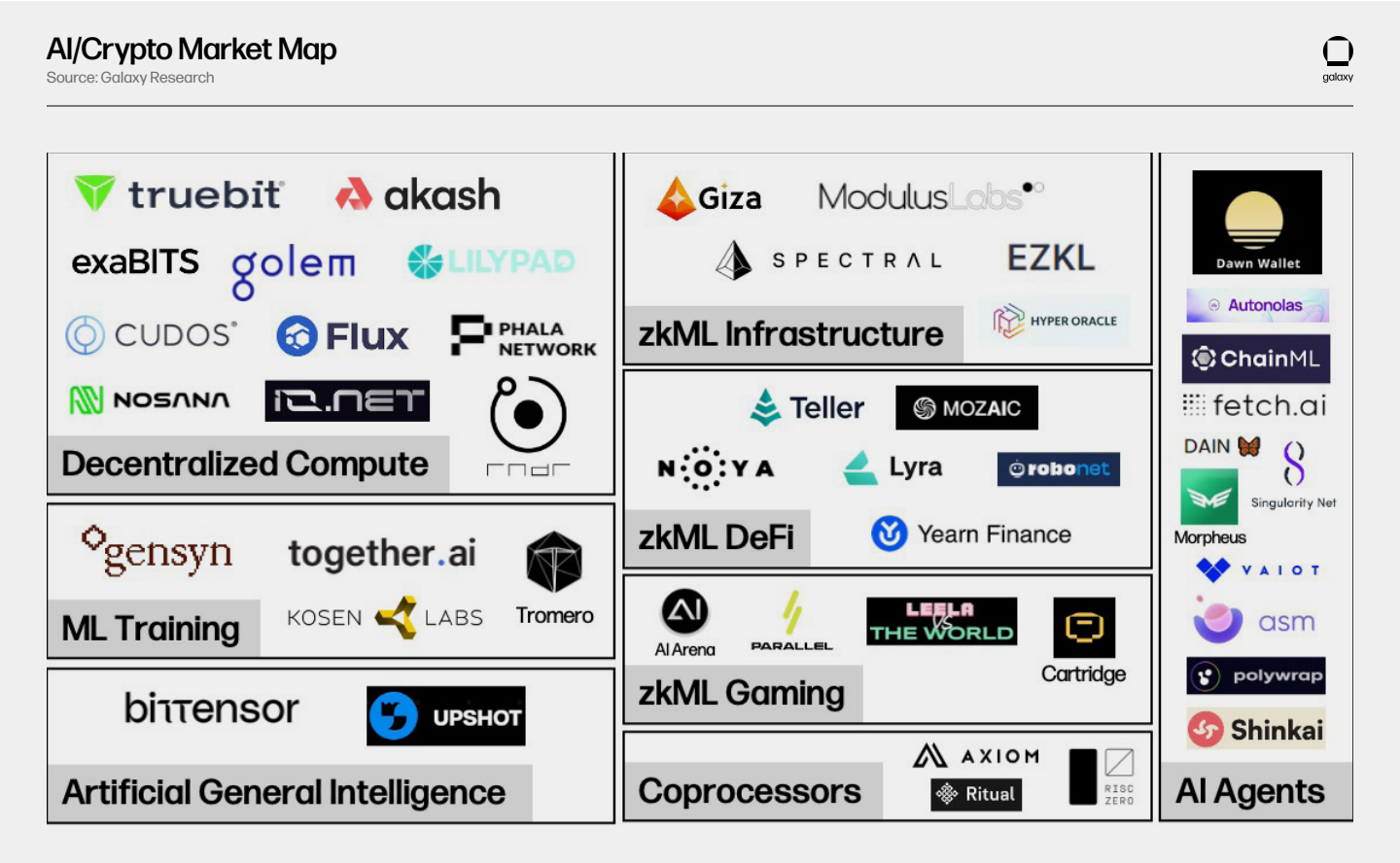
Neural Networks are one training method for AI models. They run inputs through discrete layers of algorithms, refining them until the desired output is produced. Neural networks are made up of equations that have weights which can be modified to change the output. They can require incredible amounts of data and computation to be trained so that their outputs are accurate. It is one of the most common ways that AI models are developed (ChatGPT uses a neural network process reliant on transformers).

Training is the process whereby neural networks and other AI models are developed. It requires large amounts of data to train models to correctly interpret inputs and produce accurate outputs. During the training process the weights of the model's equation are continually modified until a satisfactory output is produced. Training can be very expensive. ChatGPT, for example, uses tens of thousands of its own GPUs to process its data. Teams with fewer resources often rely on dedicated compute providers like Amazon Web Services, Azure, and Google Cloud Providers.

Inference is the actual use of an AI model to get an output or result (for example, using ChatGPT to create an outline for a paper on the intersection of crypto and AI). Inferences are used throughout the training process and in the final product. They can be expensive to run, even after training is completed, due to the computational costs, but are less computationally intensive than training.

Zero Knowledge Proofs (ZKP) allow for the verification of a claim without revealing the underlying information. This is useful in crypto for two primary reasons: 1) privacy and 2) scaling. For privacy, this enables users to transact without revealing sensitive information like how much ETH is in their wallet. For scaling, it enables off-chain computation to be proved on-chain more quickly than having to re-execute the computation. This enables blockchains and applications to run computations cheaply off-chain and then verify them on-chain. For more information on zero-knowledge and its role in the Ethereum Virtual Machine, please refer to Christine Kim's report zkEVMs: The Future of Ethereum Scalability.

AI/Crypto Market Map



Projects at the intersection of AI and crypto are still building out the underlying infrastructure needed to support on-chain AI interactions at scale.

Decentralized compute marketplaces are emerging to supply the large amounts of physical hardware, primarily in the form of graphical processing units (GPUs), needed for training and inferencing AI models. These two-sided marketplaces connect those leasing and looking to lease compute, facilitating the transfer of value and verification of compute. Within decentralized compute several subcategories are emerging that provide additional functionality. In addition to two-sided marketplaces, this report will examine machine-learning training providers that specialize in servicing verifiable training and fine-tuning outputs as well as projects working to connect compute and model generation to achieve artificial general intelligence, also frequently referred to as intelligence incentivization networks.

zkML is an emerging area of focus for projects that want to provide verifiable model outputs on-chain in a cost-effective and timely manner. These projects primarily enable applications to handle heavy compute requests offchain, and then post on-chain a verifiable output proving the offchain workload is complete and accurate. zkML is both expensive and time consuming in its current instantiation, but is increasingly being used as a solution. This is apparent in the growing number of integrations between zkML providers and DeFi/Gaming applications that want to leverage AI models.

Ample supply of compute and the ability to verify that compute on-chain opens the door for on-chain AI agents. Agents are trained models capable of executing requests on behalf of a user. Agents offer the opportunity to significantly enhance the on-chain experience, enabling users to execute complex transactions just by speaking to a chatbot. As they exist today, however, Agent projects are still focused on developing the infrastructure and tooling for easy and rapid deployment.

Decentralized Compute

Overview

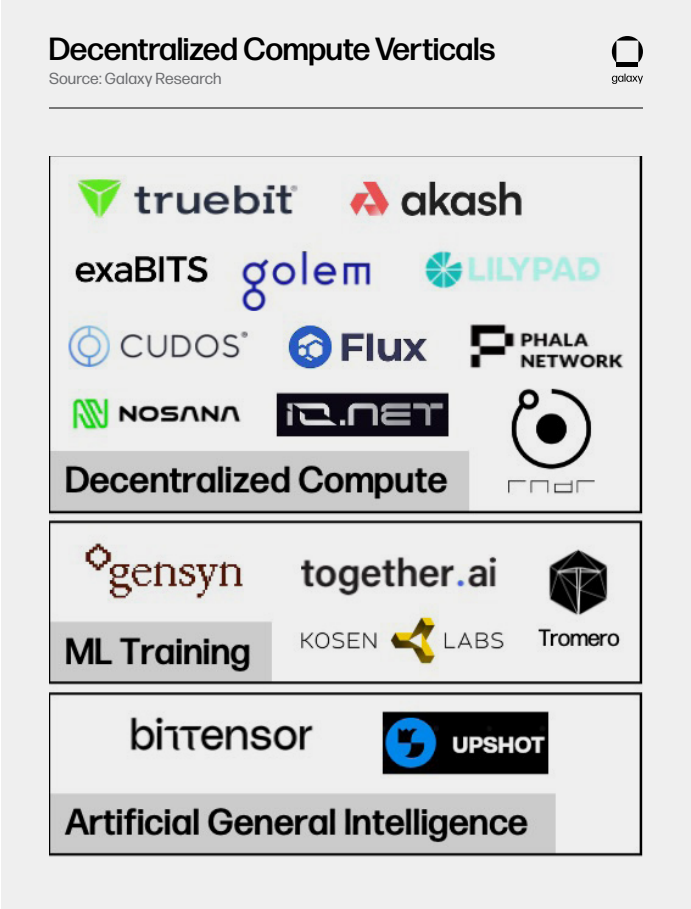
AI requires large amounts of compute, both for training models and running inferences. Over the past decade, as models have become more sophisticated, compute requirements have grown exponentially. OpenAI, for example, found that between 2012 to 2018 compute requirements for its models went from doubling every two years to every three and a half months. This has led to a surge in demand for GPUs with some crypto miners even repurposing their GPUs to provide cloud computing services (read more about this in our annual Bitcoin mining report). As competition to access compute increases and costs rise, several projects are utilizing crypto to provide decentralized compute solutions. They offer on-demand compute at competitive prices so that teams can affordably train and run models. The tradeoff, in some cases, can be performance and security.

State of the art GPUs, such as those produced by Nvidia, are in high demand. In September, Tether acquired a stake in Northern Data, a German Bitcoin miner, reportedly paying $420 million to acquire 10,000 H100 GPUs (one of the most advanced GPUs for AI training). Wait times for the best-in-class hardware can be at least six months, and in many cases longer. Making the situation even worse, companies are often required to sign long-term contracts for amounts of compute they might not even use. This can lead to situations where there is available compute, but it's not available on the market. Decentralized compute systems help address these market inefficiencies, creating a secondary market where owners of compute can sublease their excess capacity at a moment's notice, unlocking new supply.

Beyond competitive pricing and accessibility, decentralized compute's key value proposition is censorship resistance. Cutting edge AI development is increasingly dominated by large technology firms with unparalleled access to compute and data. The first key theme highlighted in the AI Index Report 2023 annual report is that industry is increasingly outpacing academia in the development of AI models, centralizing control in the hands of a few technology leaders. This has raised concerns over their ability to have an outsized influence in dictating the norms and values that underpin AI models, especially following regulatory pushes by these same tech companies to curtail AI development outside their control.

Decentralized Compute Verticals

Several models for decentralized compute have emerged in recent years, each with their own focus and tradeoffs.



Generalized Compute

Projects like Akash, io.net, iExec, Cudos, and many others are decentralized compute applications that offer access to, or will soon offer access to, specialized compute for AI training and inferences in addition to data and generalized compute solutions.

Akash is currently the only fully open source "supercloud" platform. It is a proof of stake network using Cosmos SDK. AKT, Akash's native token is used to secure the network, as a form of payment, and to incentivize participation. Akash launched its first mainnet in 2020 focused on providing a permissionless cloud compute marketplace initially featuring storage and CPU leasing services. In June 2023, Akash launched a new testnet focused on GPUs and in September launched its GPU mainnet enabling users to lease GPUs for AI training and inferences.

There are two main actors in the Akash ecosystem - Tenants and Providers. Tenants are users of the Akash network that want to buy computational resources. Providers are the compute suppliers. To match tenants and providers, Akash relies on a reverse auction process. Tenants submit their compute requirements, within which they can specify certain conditions such as the location of the servers or the type of hardware conducting the compute, and the amount they are willing to pay. Providers then submit their ask price, with the lowest bid receiving the task.

Akash validators maintain the integrity of the network. The validator set is currently limited to 100 with plans to incrementally increase over time. Anyone can become a validator by staking more AKT than the current validator with the least amount of AKT staked. AKT holders can also delegate their AKT to validators. Transaction fees and block rewards for the network are distributed in AKT. In addition, for every lease the Akash network earns a "take fee" at a rate determined by the community which is distributed to AKT holders.

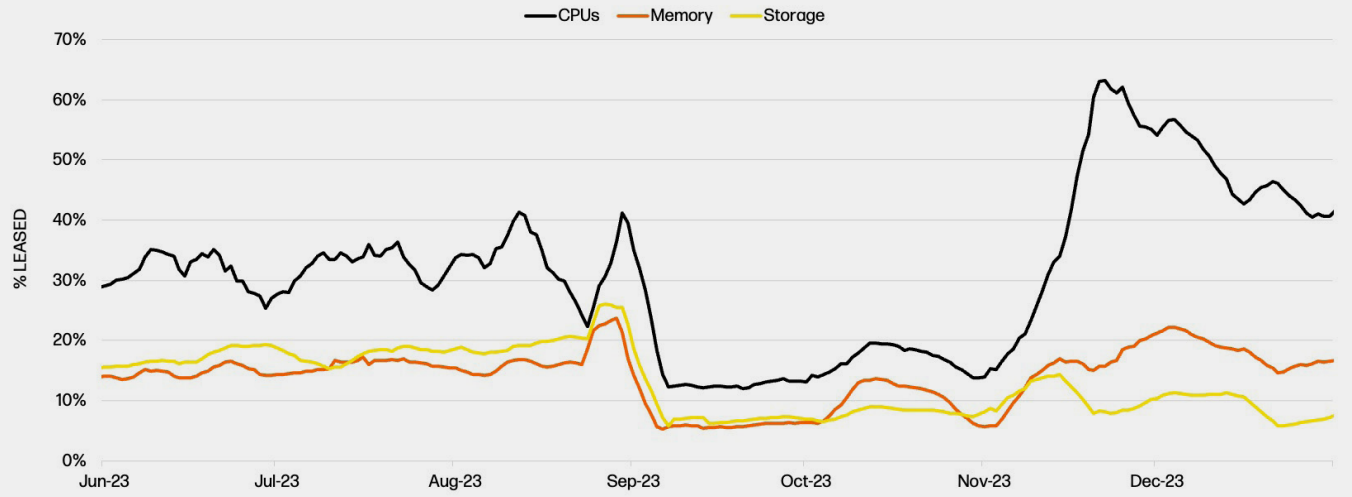
Secondary Markets

Decentralized compute marketplaces aim to fill inefficiencies in the existing compute market. Supply constraints are leading companies to hoard compute beyond what they may need, and supply is further constrained due to the structure of contracts with cloud providers which lock customers into long-term contracts even though continued access may not be required. Decentralized compute platforms unlock new supply, enabling anyone in the world with the demanded compute to become a supplier.

It remains to be seen whether surging demand for GPUs for AI training will translate to long-term network usage on Akash. Akash has long provided a marketplace for CPUs, for example, offering similar services as centralized alternatives at 70-80% discount. Lower prices, however, have not resulted in significant uptake.

Akash GPU Supply and Leases

Source: Galaxy Research



Data: Coin Metrics

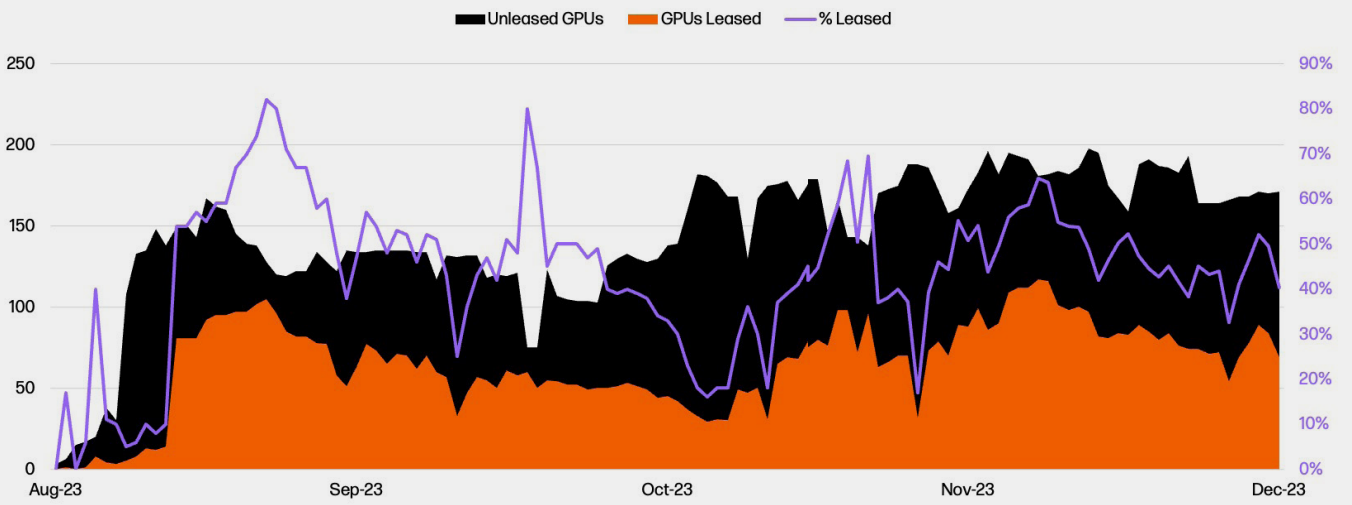
Active leases on the network have flattened out, averaging only 33% compute,16% of memory, and 13% of storage for the second of 2023. While these are impressive metrics for on-chain adoption (for reference, leading storage provider Filecoin had 12.6% storage utilization in Q3 2023), it demonstrates that supply continues to outpace demand for these products.

It has been just over half a year since Akash launched its GPU network and it's still too early to accurately gauge long-term adoption. A sign of demand, average GPU utilization to date is 44% and higher than CPUs, memory, and storage. This is primarily driven by demand for the highest quality GPUs (like A100s), with over 90% leased out.

Daily spending on Akash has also risen, nearly doubling relative to pre-GPUs. This can be partially attributed to a rise in other services used, especially CPUs, but is primarily a result of new GPU usage.

Akash GPU Supply and Leases

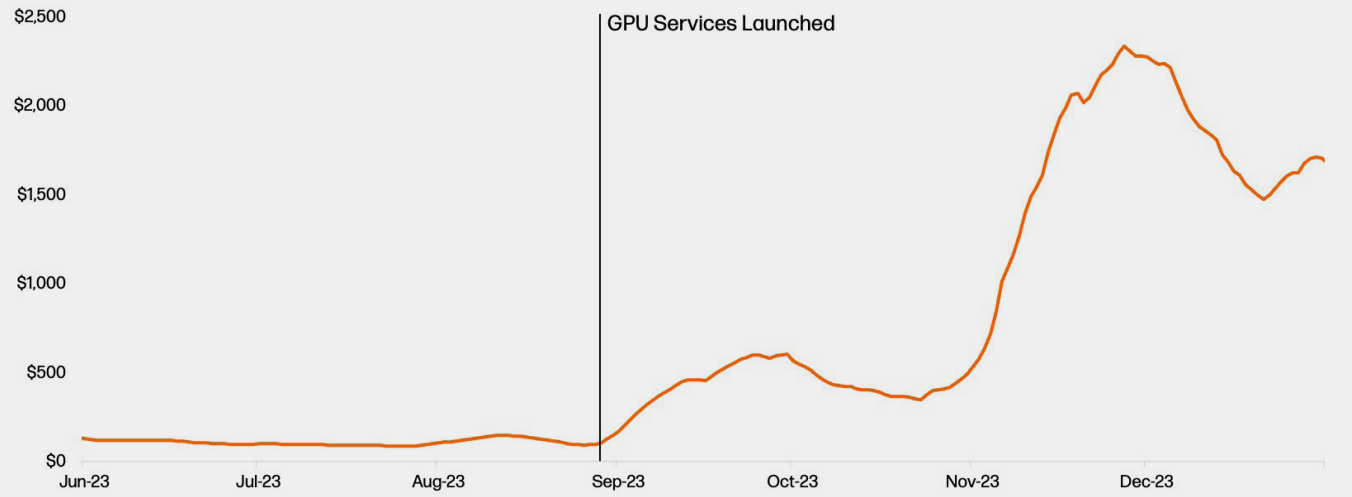
Source: Galaxy Research



Data: Cloudmos.io

Akash GPU Supply and Leases

Source: Galaxy Research



Data: Cloudmos.io

GPU Price Comparison

Source: Galaxy Research

Provider	H100 (80 GB)	A100 (80 GB)	A100 (40 GB)	RTX A600	RTX 4090	RTX 3090
Akash	$1.99	$1.50	$1.10	$0.80	$0.39	$0.30
Google Cloud	–	$5.07	$3.67	–	–	–
Amazon Web Services	–	$5.12	$4.10	–	–	–
Microsoft Azure	–	$3.67	–	–	–	–
Lamda Cloud	$1.99	$1.50	$1.10	–	–	–
CoreWeave	$4.25	$2.21	$2.06	–	–	–
Vast.AI	–	$1.35	$1.00	–	$0.35	$0.31

Data: Cloud-gpus.com, akashml.com
Note: Prices are estimates based on public information. Actual price may vary dependling on provider.

Pricing matches (or in some cases is slightly even more expensive) its centralized competitors like Lambda Cloud and Vast.ai. The incredible demand for the highest end GPUs (such as H100 and A100s) means that most owners of that equipment have little interest in listing on marketplaces where they face competitive pricing.

While initial interest is promising, there remain barriers to adoption (discussed further below). Decentralized compute networks will need to do more to generate both demand and supply and teams are experimenting with how best to attract new users. In the beginning of 2024, for example, Akash passed Proposal 240 to increase AKT emissions for GPU suppliers and incentivize more supply, specifically targeting higher-end GPUs. Teams are also working on rolling out proof-of-concept models to demonstrate to prospective users the real-time capabilities of their networks. Akash is training their own foundational model and has already launched chatbot and image generation offerings that create outputs using Akash GPUs. Similarly, io.net has developed a stable diffusion model and is rolling out new network functionalities which better mimic the performance and scale of traditional GPU datacenters.

Decentralized Machine Learning Training

In addition to generalized compute platforms that can service AI needs, a set of specialized AI GPU providers focused on machine learning model training are also emerging. Gensyn, for example, is "coordinating electricity and hardware to build collective intelligence" with the view that, "If someone wants to train something, and someone is willing to train it, then that training should be allowed to happen."

The protocol has four primary actors: submitters, solvers, verifiers, and whistleblowers. Submitters submit tasks to the network with training requests. These tasks include the training objective, the model to be trained, and training data. As part of the submission process, submitters pay a fee up-front for the estimated compute required from the solver.

Once submitted, tasks are assigned to solvers who conduct the actual training of the models. Solvers then submit completed tasks to verifiers who are responsible for checking the training to ensure it was done correctly. Whistleblowers are responsible for ensuring that verifiers behave honestly. To incentivize whistleblowers to participate in the network, Gensyn plans to periodically provide purposefully incorrect proofs that reward whistleblowers for catching them.

Beyond providing compute for AI-related workloads, Gensyn's key value proposition is its verification system, which is still in development. Verification is necessary to ensure that external computations by GPU providers are performed correctly (i.e., to ensure that a user's model is trained the way they want it to be). Gensyn tackles this problem with a unique approach, leveraging novel verification methods called, "Probabilistic proof-of-learning, Graph-based pinpoint protocol, and Truebit-style incentive games." This is an optimistic solving mode that allows a verifier to confirm that a solver has correctly run a model without having to completely rerun it themselves, which is a costly and inefficient process.

In addition to its innovative verification method, Gensyn also claims to be cost effective relative to centralized alternatives and crypto competitors - providing ML training at up to 80% cheaper than AWS while outcompeting similar projects like Truebit in testing.

Training Price Comparison

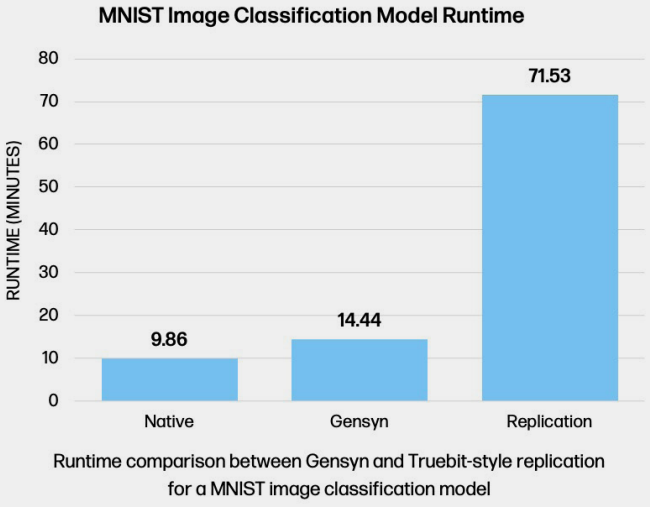
Source: Galaxy Research

Provider	Approximate Hourly Cost for ML Training Work (V100-equivalent)	Scalability
Ethereum	$15,700	Low
Truebit (+Ethereum)	$12.00	Low
GCP on-demand	$2.50	Medium
AWS on-demand	$2.00	Medium
Golem Network	$1.20	Low
Vast.ai	$1.20	Low
AWS spot instances (unreliable)	$0.90	Medium
GCP spot instances (unreliable)	$0.75	Medium
Gensyn (projected)	$0.40	High
Single GPU in datacentre	$0.40	None
Single person GPU	$0.28	None

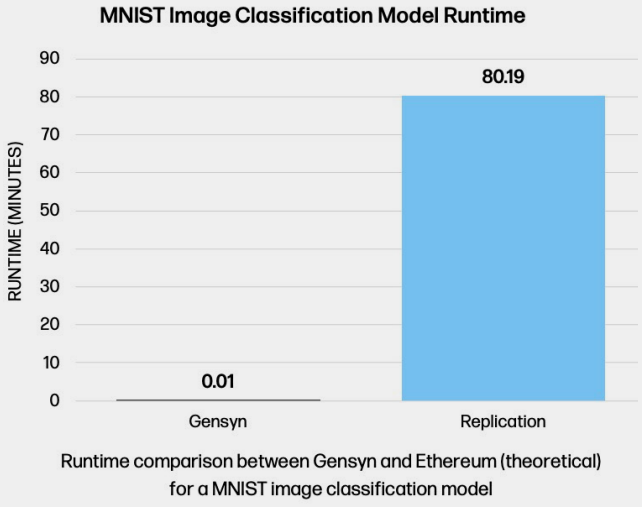
Data: https://docs.gensyn.ai/litepaper/

Training Performance Comparison

Source: Galaxy Research



Runtime comparison between Gensyn and Truebit-style replication for a MNIST image classification model



Runtime comparison between Gensyn and Ethereum (theoretical) for a MNIST image classification model

https://docs.gensyn.ai/litepaper/

Whether these initial results can be replicated at scale across a decentralized network remains to be seen. Gensyn wants to harness excess compute from providers like small data centers, retail users, and in the future even smaller mobile devices like cell phones. However, as the Gensyn team itself has admitted, relying on heterogenous compute providers introduces several new challenges.

For centralized providers like Google Cloud Providers and Coreweave, compute is expensive while communication between that compute (bandwidth and latency) is cheap. These systems are designed to enable communication between hardware as quickly as possible. Gensyn flips that framework on its head, reducing compute costs by enabling anyone in the world to supply GPUs but increasing communication costs as the network must now coordinate compute jobs across heterogeneous hardware located far apart. Gensyn has not yet launched, but is a proof of concept of what might be possible when it comes to building decentralized machine learning training protocols.

Decentralized Generalized Intelligence

Decentralized compute platforms are also opening the design possibilities for artificial intelligence creation methods. Bittensor is a decentralized compute protocol built on Substrate that is trying to answer the question of, "how can we turn AI into a collaborative approach?" Bittensor aims to decentralize and commodify artificial intelligence generation. Launched in 2021, the protocol wants to harness the power of collaborative machine learning models to continually iterate and produce better artificial intelligence.

Bittensor draws inspiration from Bitcoin, with a twenty-one million supply of its native currency TAO and a four-year halving cycle (the first halving will be in 2025). Rather than use Proof of Work to generate the correct nonce and earn a block reward, Bittensor relies on "Proof of Intelligence," requiring miners to run models that produce outputs in response to inference requests

Incentivizing Intelligence

Bittensor originally relied on a Mixture of Experts (MoE) model to produce outputs. When inference requests are submitted, rather than relying on one generalized model, MoE models relay the inference request to the most accurate models for a given input type. Think of building a house where you hire a variety of specialists for different aspects of the construction process (ex: architects, engineers, painters, construction workers etc...). MoE applies this to machine learning models, attempting to harness the outputs of different models depending on the input. As Bittensor founder Ala Shaabana explained, it's like "speaking to a room of smart people and getting the best answer rather than speaking to one person." Due to challenges with ensuring the proper routing, synchronization of messages to the correct model, and incentivization, this approach has been sidelined until the project is more developed.

There are two primary actors in Bittensor network: validators and miners. Validators are tasked with sending inference requests to miners, reviewing their outputs, and ranking them based on the quality of their responses. To ensure their rankings are reliable, validators are given "vtrust" scores based on how well their rankings align with the rankings of other validators. The higher a validator's vtrust score, the more TAO emissions they earn. This is meant to incentivize validators to reach consensus on model rankings over time, as the more validators that reach agreement on rankings the higher their individual vtrust scores.

Miners, also called servers, are network participants that run the actual machine learning models. Miners compete against each other to provide validators with the most accurate outputs for a given query, earning more TAO emissions the more accurate their output. Miners can generate those outputs however they want. For example, it is entirely possible in a future scenario that a Bittensor miner could have previously trained models on Gensyn that they use to earn TAO emissions.

Today, most interactions happen directly between validators and miners. Validators submit inputs to miners and request outputs (i.e., training the model). Once a validator has queried the miners on the network and received their responses, they then rank the miners and submit their rankings to the network.

This interaction between validators (who rely on PoS) and miners (who rely on Proof of Model, a form of PoW) –is called Yuma Consensus. It seeks to incentivize miners to produce the best outputs to earn TAO emissions and validators to accurately rank miner outputs to earn a higher vtrust score and increase their TAO rewards forming the network's consensus mechanism.

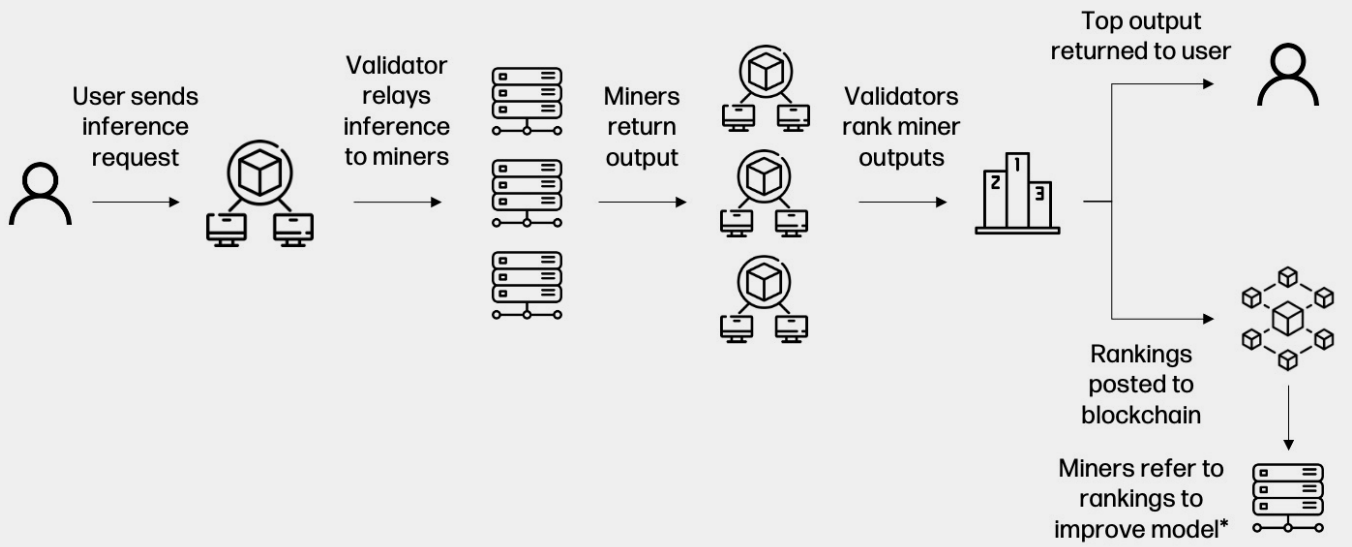
Subnets and Applications

Interactions on Bittensor primarily consist of validators submitting requests to miners and evaluating their outputs. As the quality of the contributing miners increases and the overall intelligence of the network grows, however, Bittensor will create an application layer on top of its existing stack so that developers can build applications that query the Bittensor network.

In October 2023, Bittensor completed an important step toward achieving this with the introduction of subnets through its Revolution upgrade. Subnets are individual networks on Bittensor that incentivize specific behaviors. Revolution opens the network to anyone interested in creating a subnet. In the months since its release, over 32 subnets have been launched, including ones for text prompting, data scraping, image generation, and storage. As subnets mature and become product-ready, subnet creators will also create application integrations, enabling teams to build applications that query a specific subnet. Some applications (chatbot, image generator, twitter reply bot, prediction market) do exist today, but there are no formal incentives for validators to accept and relay those queries beyond grants from the Bittensor foundation.

To provide a clearer illustration, here is an example of how Bittensor might work once applications are integrated into the network.

Bittensor Inference Process
Source: Galaxy Research



*Not applicable to all subnets

Subnets earn TAO based on their performance evaluated by the root network. The root network sits on top of all the subnets, essentially acting as a special kind of subnet, and is managed by the 64 largest subnet validators by stake. Root network validators rank subnets based on their performance and distribute TAO emissions to subnets periodically. In this way, individual subnets act as the miners for the root network.

Bittensor Outlook

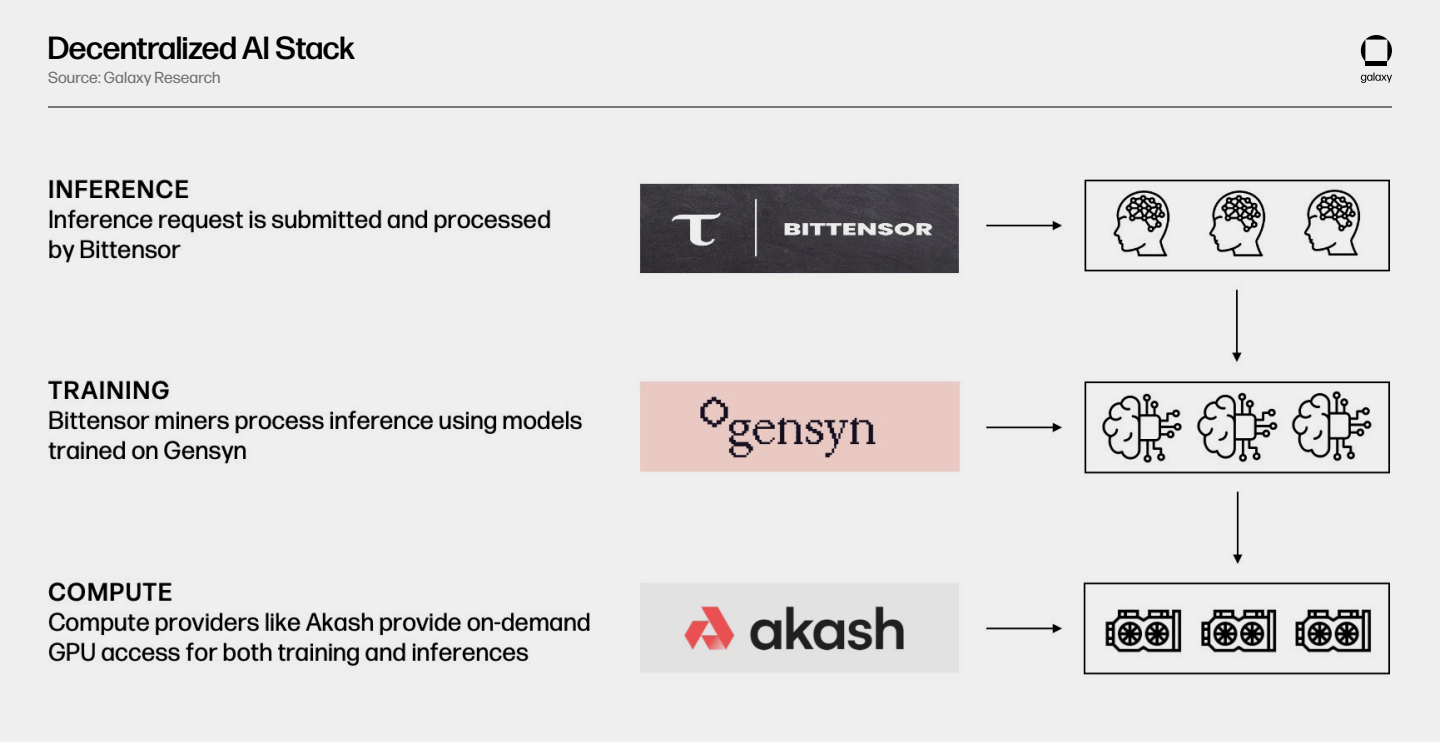
Bittensor is still experiencing growing pains as it expands the protocol's functionality to incentivize intelligence generation across multiple subnets. Miners continue to devise new ways to attack the network to earn more TAO rewards, for example by slightly modifying the output of a highly rated inference run by their model and then submitting multiple variations. Governance proposals impacting the whole of the network can only be submitted and implemented by the Triumvirate, which is composed entirely of Opentensor Foundation stakeholders (important to note that proposals require approval by the Bittensor Senate composed of Bittensor validators prior to implementation). And the project's tokenomics are under the process of being revamped to improve incentives for TAO's usage across subnets. The project is also rapidly gaining notoriety for its unique approach, with the CEO of one of the most popular AI websites HuggingFace indicating Bittensor should add its resources to the website.

In a recently published piece by a core dev called "Bittensor Paradigm," the team lays out their vision for Bittensor to eventually evolve to be "agnostic to what is being measured." In theory, this could enable Bittensor to develop subnets incentivizing any type of behavior all powered by TAO. There remain considerable practical constraints – most notably demonstrating these networks are capable of scaling to handle such a diverse set of processes and that the underlying incentives drive progress that outpaces centralized offerings.

Building a Decentralized Compute Stack for AI Models

The above sections provide a high-level overview of the various types of decentralized artificial intelligence compute protocols being developed. While early in their development and adoption, they provide the foundation of an ecosystem that could eventually facilitate the creation of "AI building blocks," like DeFi's "Money Legos" concept. The composability of permissionless blockchains opens the possibility for each protocol to build on top of the other to provide a more comprehensive decentralized artificial intelligence ecosystem.

For example, here is one way in which Akash, Gensyn, and Bittensor might all interact to respond to an inference request.



To be clear, this is simply an example of what might be possible in the future, not a representation of the current ecosystem, existing partnerships, or likely outcomes. Constraints on interoperability, as well as other considerations described below, considerably limit integration possibilities today. Beyond that, fragmentation of liquidity and the need for using multiple tokens can be detrimental to the user experience, something that has been pointed out by founders of both Akash and Bittensor.

Other Decentralized Offerings

In addition to compute, several other decentralized infrastructure services are being rolled out to support crypto's emerging AI ecosystem. To list them all is beyond the scope of this report, but a few interesting and illustrative examples include:

- Ocean: A decentralized data marketplace. Users can create data NFTs representative of their data that can be purchased using data tokens. Users can both monetize their data and have greater sovereignty over it while providing teams working on AI with access to the data needed to develop and train models.
- Grass: A decentralized bandwidth marketplace. Users can sell their excess bandwidth to AI companies who use it to scrape data from the internet. Built on the Wynd Network, this not only enables individuals to monetize their bandwidth, but also provides purchasers of the bandwidth with a more diverse set of viewpoints into what individual users see online (as an individual's internet access is usually tailored to their IP address specifically).
- HiveMapper: Building out a decentralized maps offering comprised of information collected from everyday automobile drivers. HiveMapper relies on AI to interpret the images collected from user's dashboard cameras and rewards users with tokens for helping to fine-tune the AI model through reinforced human learning feedback (RHLF).

Collectively, these point toward the near endless opportunities to explore decentralized marketplace models that support AI models, or the surrounding infrastructure needed to develop them. For now, these projects are mostly in the proof-of-concept stage and much more research and development is needed to demonstrate they can operate at the scale needed to provide comprehensive AI services.

Outlook

Decentralized compute offerings are still in the early stages of development. They are just beginning to roll out access to state-of-the-art compute capable of training the most powerful AI models in production. For them to gain meaningful market share, they'll need to demonstrate practical advantages compared to centralized alternatives. Potential triggers for broader adoption include:

- **GPU Supply/Demand.** GPU scarcity coupled with rapidly increasing compute demand is leading to a GPU arms race. OpenAI has already once limited access to its platform due to GPU constraints. Platforms like Akash and Gensyn can provide cost competitive alternatives for teams in need of high-performance compute. The next 6-12 months is an especially unique opportunity for decentralized compute providers to onboard new users that are forced to consider decentralized offerings given lack of access in the broader market. Coupled with increasingly performant open-source models like Meta's LLaMA 2, users no longer face the same barriers to deploying effective fine-tuned models, making compute resources the primary bottleneck. The existence of the platforms themselves, however, does not ensure an adequate supply of compute and corresponding demand from consumers. Sourcing high-end GPUs remains difficult, and cost is not always the primary motivation on the demand side. These platforms will be challenged to demonstrate the practical benefit of using a decentralized compute option – whether that be due to cost, censorship resistance, uptime and resilience, or accessibility – to accumulate sticky users. They will have to move fast. GPU infrastructure investment and buildout is happening at extraordinary rates.

- **Regulation**. Regulation continues to be a headwind to the decentralized compute movement. In the near term, lack of clear regulation means that both providers and users face potential risks for using these services. What if a supplier provides compute or a buyer purchases compute from a sanctioned entity unknowingly? Users may be hesitant to use a decentralized platform that lacks the controls and oversight of a centralized entity. Protocols have tried to mitigate these concerns by incorporating controls into their platforms or adding filters to access known compute providers only (i.e., have provided know-your-customer (KYC) information), but more robust methods that protect privacy while ensuring compliance will be needed for adoption. In the short term, we are likely to see the emergence of KYC and regulatory compliant platforms that limit access to their protocols to address these concerns. Additionally, discussions around possible new US regulatory frameworks, best illustrated by the release of the *Executive Order on the Safe, secure, and Trustworthy Development and Use of Artificial Intelligence*, highlight the potential for regulatory action that further curtails access to GPUs.

- **Censorship**. Regulation goes both ways and decentralized compute offerings could benefit from actions taken to limit access to AI. In addition to the Executive Order, OpenAI founder Sam Altman has testified in Congress on the need for regulatory agencies that issue licenses for AI development. Discussion around AI regulation is just getting started, but any such attempts to curtail access or censor what can be done with AI may accelerate adoption of decentralized platforms that have no such barriers. The November OpenAI leadership shakeup (or lack thereof) further demonstrates the risks of empowering decision making for the most powerful existing AI model to just a few. Furthermore, all AI models necessarily reflect the biases of those who created them, whether intentionally or not. One way to eliminate those biases is to make models as open as possible to fine-tuning and training, ensuring that models of all varieties and biases can always be accessed by anyone anywhere.

- **Data Privacy.** When integrated with external data and privacy solutions that provide users with autonomy over their data, decentralized compute may become more attractive than centralized alternatives. Samsung fell victim to this when they realized that engineers were using ChatGPT to help with chip design and leaking sensitive information to ChatGPT. Phala Network and iExec claim to offer users SGX secure enclaves to protect user data and ongoing research in fully homomorphic encryption could further unlock privacy ensured decentralized compute. As AI becomes further integrated into our lives, users will place a greater premium on being able to run models on applications that have privacy baked into them. Users will also demand services that enable data composability so that they can seamlessly port their data from one model to another.

- User Experience (UX). UX continues to be a significant barrier to broader adoption of all types of crypto applications and infrastructure. This is no different for decentralized compute offerings, and in some cases is exacerbated by the need for developers to understand both crypto and AI. Improvements are needed from the basics such as onboarding and abstracting away interaction with the blockchain to providing the same high quality outputs as current market leaders. This is glaringly evident given the fact that many operational decentralized compute protocols providing cheaper offerings struggle to gain regular usage.

Smart Contracts & zkML

Smart contracts are a core building block of any blockchain ecosystem. Given a specific set of conditions, they execute automatically and reduce or remove the need for a trusted third party, enabling the creation of complex decentralized applications like those seen in DeFi. As they exist for the most part today, however, smart contracts are still limited in their functionality in that they execute based on preset parameters that must be updated.

For example, a lend/borrow protocol smart contract is deployed with specifications for when to liquidate a position based on certain loan to value ratios. While useful in a static environment, in a dynamic situation where risk is constantly shifting, these smart contracts must be continually updated to account for changes in risk tolerance, creating challenges for contracts that are not governed through centralized processes. DAOs, for example, that are reliant on decentralized governance processes, may not be able to react quickly enough to respond to systemic risks.

Smart contracts that integrate AI (i.e., machine learning models) are one possible way to enhance functionality, security, and efficiency while improving the overall user experience. These integrations also introduce additional risks, though, as it is impossible to ensure that the models underpinning these smart contracts cannot be exploited or account for longtail situations (which are notoriously hard to train models on given the scarcity of data inputs for them).

Zero Knowledge Machine Learning (zkML)

Machine learning requires large amounts of compute to run complex models, which prevents AI models from being directly run inside smart contracts due to high costs. A DeFi protocol providing users access to a yield optimizing model, for example, would struggle to run that model on-chain without having to pay prohibitively high gas fees. One solution is to increase the computational power of the underlying blockchain. However, this also increases the demands on the chain's validator set, potentially undermining decentralization properties. Instead, some projects are exploring the use of zkML for verifying outputs in a trustless manner without needing intensive on-chain computation.

One commonly shared example that illustrates the usefulness of zkML is when a user needs someone else to run data through a model and also verify that their counterparty actually ran the correct model. Maybe a developer is using a decentralized compute provider to train their models and worries that the provider is trying to cut costs by using a cheaper model with a near unnoticeable difference in output. zkML enables the compute provider to run the data through their models and then generate a proof that can be verified on-chain to prove the model's output for the given input is correct. In this case, the model provider would have the added advantage of being able to offer their models without having to reveal the underlying weights that produce the output.

The opposite could also be done. If a user wants to run a model using their data but does not want the project providing the model to have access to their data due to privacy concerns (i.e., in the case of a medical exam or proprietary business information), then the user could run the model on their data without sharing it and then verify they ran the correct model with a proof. These possibilities considerably expand the design space for the integration of AI and smart contract functionality by tackling prohibitive compute restrictions.

Infrastructure and Tooling

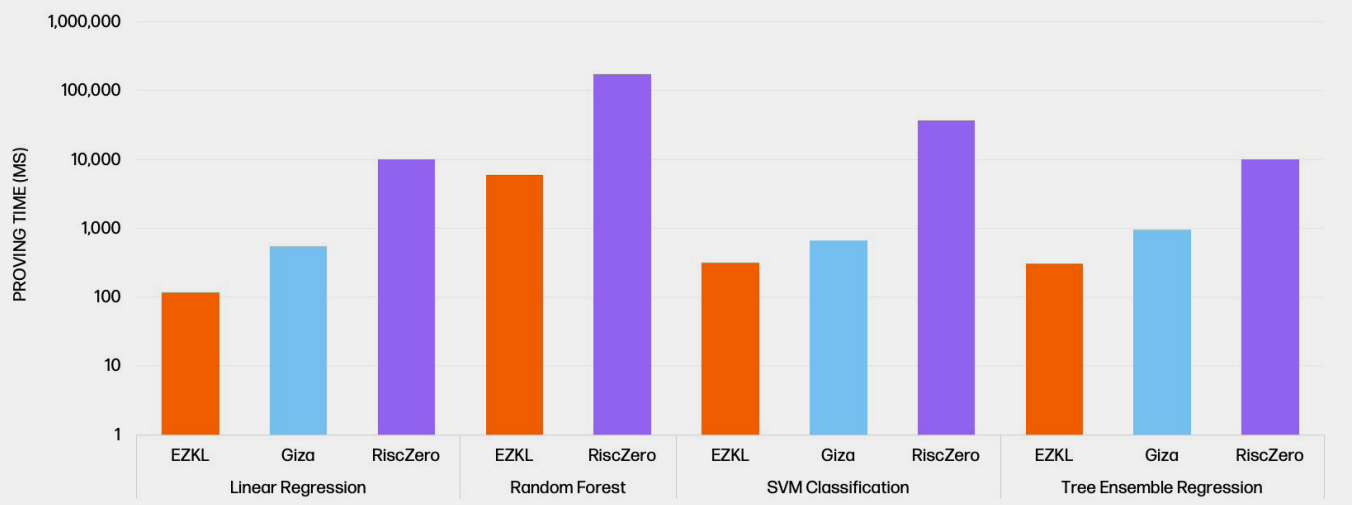
Given the early state of the zkML space, development is primarily focused on building out the infrastructure and tooling needed for teams to convert their models and outputs into proofs that can be verified on-chain. These products abstract away the zero-knowledge aspect of development as much as possible.

EZKL and Giza are two projects building this tooling by providing verifiable proofs of machine learning model execution. Both help teams build machine learning models to ensure that those models can then be executed in a way that the results can be trustlessly verified on-chain. Both projects use the Open Neural Network Exchange (ONNX) to transform machine learning models written in common languages like TensorFlow and Pytorch into a standard format. They then output versions of those models that also produce zk-proofs when executed. EZKL is open source and produces zk-SNARKS while Giza is closed source and produces zk-STARKS. Both projects are currently only EVM compatible.

EZKL has demonstrated significant progress over the past few months in enhancing their zkML solution, primarily focused on reducing costs, improving security, and speeding up proof generation. In November 2023, for example, EZKL integrated a new open-source GPU library that reduces aggregate proof time by 35% and in January EZKL announced Lilith, a software solution for integrating high-performance compute clusters and orchestrating concurrent jobs when using the EZKL proving system. Giza is unique in that in addition to providing tooling for creating verifiable machine learning models, they also plan to implement a web3 equivalent of Hugging Face, opening up an user marketplace for zkML collaboration and model sharing as well as eventually integrating decentralized compute offerings. In January EZKL released a benchmark assessment comparing EZKL, Giza, and RiscZero (discussed below) performance. EZKL demonstrated faster proving times and memory usage.

zkML Proving Time

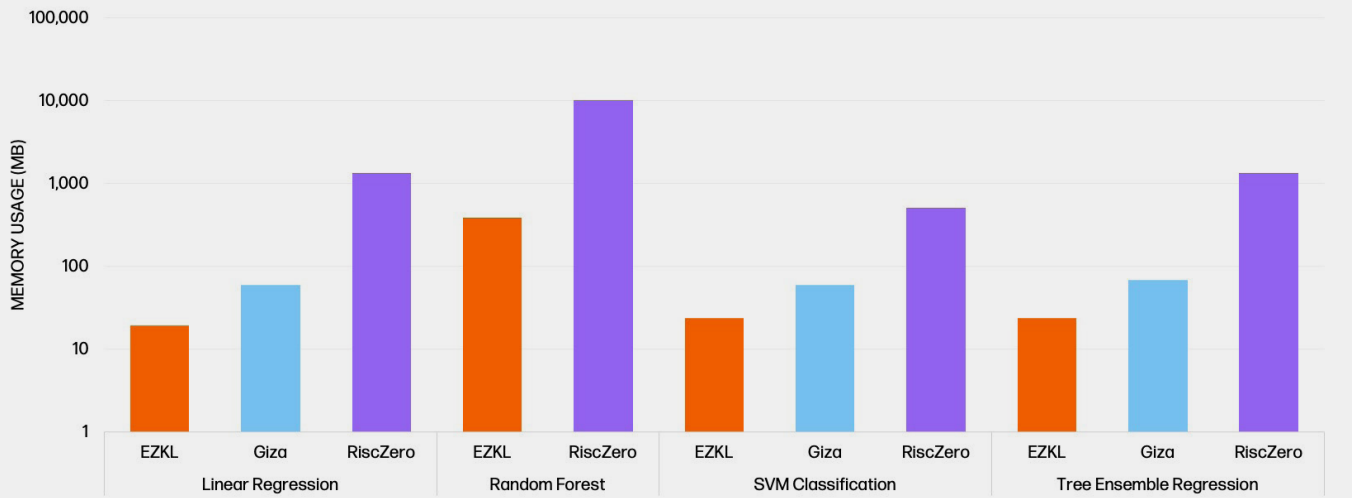
Source: Galaxy Research



Data: https://blog.ezkl.xyz/post/benchmarks/

zkML Memory Usage

Source: Galaxy Research



Data: https://blog.ezkl.xyz/post/benchmarks/

Modulus Labs is also developing a new zk-proof technique custom tailored for AI models. Modulus published a paper called The Cost of Intelligence (hinting at the incredibly high costs to run AI models on-chain), that benchmarked existing zk-proof systems at the time to identify capabilities and bottlenecks for improving AI model zk-proofs. Published in January 2023, the paper demonstrates that existing offerings are simply too expensive and inefficient to enable AI applications at scale. Building on their initial research, in November Modulus introduced Remainder, a specialized zero-knowledge prover built specifically to reduce costs and proving time for AI models with a goal to make it economically feasible for projects to integrate models into their smart contracts at scale. Their work is closed-source and so could not be benchmarked to the above solutions, but was recently referenced in Vitalik's blog post on crypto and AI.

Tooling and infrastructure development is critical to the future growth of the zkML space because it significantly reduces friction for teams that need to deploy zk circuits necessary to run verifiable offchain computation. The creation of secure interfaces that enable non-crypto native builders working in machine learning to bring their models on-chain will enable greater experimentation of applications with truly novel use cases. Tooling also addresses a major obstacle to broader zkML adoption, a lack of developers knowledgeable and interested in working at the intersection of zero-knowledge, machine learning, and cryptography.

Coprocessors

Additional solutions in development, referred to as "coprocessors," include RiscZero, Axiom, and Ritual. The term coprocessor is mostly semantics – these networks fulfill many different roles, including verifying offchain compute on-chain. As with EZKL, Giza, and Modulus, they aim to completely abstract the zero-knowledge proof generation process, creating essentially zero-knowledge virtual machines capable of executing programs offchain and generating proofs for verification on-chain. RiscZero and Axiom can service simple AI models as they are meant to be more general-purpose coprocessors while Ritual is purpose-built for use with AI models.

Infernet is the first instantiation of Ritual and includes an *Infernet SDK* that allows developers to submit inference requests to the network and receive an output and proof (optionally) in return. An *Infernet Node* receives these requests and handles the computation off-chain before returning an output. For example, a DAO could create a process for ensuring all new governance proposals meet certain preconditions prior to being submitted. Every time a new proposal is submitted, the governance contract triggers an inference request through Infernet calling a DAO-specific governance trained AI model. The model reviews the proposal to ensure all necessary criteria was submitted and returns the output and proof, either approving or denying the proposal's submission.

Over the coming year, the Ritual team plans to roll out additional features that form a base infrastructure layer called the Ritual Superchain. Many of the projects discussed previously could plug into Ritual as services providers. Already, the Ritual team has integrated with EZKL for proof generation and will likely soon add functionality from others leading providers. Infernet nodes on Ritual could also use Akash or io.net GPUs and query models trained on Bittensor subnets. Their end goal is to be the go-to provider for open AI infrastructure, capable of servicing machine-learning and other AI-related tasks from any network across any workload.

Applications

zkML helps reconcile the contradiction between blockchains and AI, where the former are inherently resource constrained and the latter requires large amounts of compute and data. As one of the founders of Giza put it, "The use cases are so abundant... it's kind of like asking in Ethereum's early days what are the use cases of smart contracts...what we are doing is just expanding the use cases of smart contracts." As highlighted above, however, development today is primarily taking place at the tooling and infrastructure level. Applications are still in the exploratory phase, with teams challenged to demonstrate that the value generated from implementing models using zkML outweighs the complexity and costs of doing it.

Some applications today include:

- **Decentralized Finance**. zkML upgrades the design space for DeFi by enhancing smart contract capabilities. DeFi protocols provide machine learning models with large amounts of verifiable and immutable data that can be used to generate yield generating or trading strategies, risk analysis, UX, and much more. Giza, for example, has a partnership with Yearn Finance to build a proof-of-concept automated risk assessment engine for Yearn's new v3 vaults. Modulus Labs has worked with Lyra Finance on incorporating machine learning into their AMMs, is partnered with Ion Protocol to implement a model for analyzing validator risk, and is helping Upshot verify their AI-powered NFT price feeds. Protocols like NOYA (which is leveraging EZKL) and Mozaic are providing access to proprietary offchain models that give users access to automated yield farming while enabling them to verify the data inputs and proofs on-chain. Spectral Finance is building on-chain credit scoring engines to predict the likelihood of Compound or Aave borrowers defaulting on their loans. These so-called "De-Ai-Fi" products are likely to become much more prevalent in the years ahead thanks to zkML.

- **Gaming**. Gaming has long been considered ripe for disruption and enhancement by public blockchains (for more on crypto and gaming refer to this Galaxy Digital report - The History of Gaming and its Web3 Future). zkML makes on-chain gaming with artificial intelligence possible. Modulus Labs has already implemented

proof of concepts for simple on-chain games. Leela vs the World is a game theoretic chess game in which users face off against an AI chess model, with zkML verifying that every move Leela makes is based on the model the game says it is running. Similarly, teams have used EZKL frameworks to build simple singing contests and on-chain tictactoe. Cartridge is using Giza to enable teams to deploy fully on-chain games, most recently highlighting a simple AI driving game where users compete to create better models for a car trying to avoid obstacles. While simple, these proof-of-concept points toward future implementations enabling more complex on-chain verifications like sophisticated NPC actors capable of interacting with in-game economies as seen in AI Arena, a super smash brothers like game where players train their fighters which are then deployed as AI models to fight.

- **Identity, Provenance, and Privacy**. Crypto is already being used as a means of verifying authenticity and combatting increasing amount of AI generated/manipulated content and deep fakes. zkML can advance those efforts. WorldCoin is a proof of personhood solution that requires users to scan their iris to generate a unique ID. In the future, biometric IDs could be self-custodied on personal devices using encrypted storage with the models needed to verify those biometrics being run locally. Users could then provide proof of their biometrics without needing to reveal who they are, combatting sybil attacks while ensuring privacy. This can also be applied to other inferences requiring privacy, such as using models to analyze medical data/images to detect diseases, verifying personhood and developing matching algorithms in dating applications, or for insurance and loan agencies that need to verify financial information.

Outlook

zkML is still in the experimental stage with most projects focused on building out infrastructure primatives and proof of concepts. Challenges today include computational costs, memory limitations, model complexity, limited tooling and infrastructure, and developer talent. Simply put, there is considerably more work to do before zkML can be implemented at a scale needed for consumer products.

As the field matures, however, and these limitations are addressed, zkML will become a critical component of AI and crypto integration. At its core, zkML promises the ability to bring off-chain compute of any size on-chain while maintaining the same or close to the same security assurances as if the computation had been run on-chain. Until that vision is realized, however, early users of the technology will continue to have to balance the tradeoffs between the privacy and security of zkML and the efficiency of alternatives.

AI Agents

One of the more exciting integrations of AI and crypto is ongoing experimentation with AI Agents. Agents are autonomous bots capable of receiving, interpreting, and executing tasks using an AI model. This could be anything from having an always available personal assistant that is fine-tuned to your preferences to hiring a financial agent that manages and adjusts your portfolio according to your risk preferences.

Agents and crypto fit well together because of the permissionless and trustless payments infrastructure crypto provides. Once trained, agents can be given a wallet so they can transact with smart contracts on their own. Simple agents today, for example, can crawl the internet for information and then make trades on prediction markets based on a model.

Agent Providers

Morpheus is one of the newest open-source agent projects coming to market on Ethereum and Arbitrum in 2024. Its white paper was anonymously published in September 2023, providing the foundation for a community to form and build around (including notable figures like Erik Vorhees). The white paper includes a downloadable Smart Agent Protocol, which is an open source LLM that can be run locally, managed by a user's wallet, and interact with smart contracts. It uses a Smart Contract Rank to help the agent determine which smart contracts are safe to interact with based on criteria such as number of transactions processed.

The white paper also provides a framework for building out the Morpheus Network, such as the incentive structures and infrastructure needed to make the Smart Agent Protocol operational. This includes incentivizing contributors to build out frontends for interacting with the agents, APIs for developers to build applications that can plug into the agents so they can interact with each other, and cloud solutions for enabling users to access compute and storage needed to run an agent on an edge device. Initial funding for the project launched in early February with the full protocol expected to launch in 2Q24.

Decentralized Autonomous Infrastructure Network (DAIN) is a new agent infrastructure protocol building out an agent-to-agent economy on Solana. DAIN aims to make it so that agents from different businesses can seamlessly interact with each other via a universal API, considerably opening the design space for AI agents with a focus on implementing agents capable of interacting with both web2 and web3 products. In January, DAIN announced their first partnership with Asset Shield enabling users to add "agent signers" to their multisig that are capable of interpreting transactions and approving/denying based on rules set by the user.

Fetch.AI was one of the first AI Agent protocols deployed and has developed an ecosystem for building, deploying, and using Agents on-chain using its FET token and Fetch.AI wallet. The protocol provides a comprehensive suite of tools and applications for using Agents, including in-wallet functionality for interacting with and ordering agents.

Autonolas, whose founders include a previous member from the Fetch team, is an open marketplace for the creation and usage of decentralized AI agents. Autonolas also provides a set of tooling for developers to build AI agents that are hosted offchain and can plug into multiple blockchains including Polygon, Ethereum, Gnosis Chain, and Solana. They currently have a few active agent proof of concept products including for use in prediction markets and DAO governance.

SingularityNet is building out a decentralized marketplace for AI agents where people can deploy narrowly focused AI agents that can be hired by other people or agents to execute complex tasks. Others, like AlteredStateMachine, are building out AI Agent integrations with NFTs. Users mint NFTs with randomized attributes that give them strengths and weaknesses for different tasks. These agents can then be trained to enhance certain attributes for uses like gaming, DeFi, or as a virtual assistant and traded with other users.

Collectively, these projects envision a future ecosystem of agents capable of working together to not only execute tasks but help build artificial general intelligence. Truly sophisticated agents will have the ability to fulfill any user task autonomously. For example, rather than having to ensure an Agent already has integrated with an external API (like a travel booking website) before using it, fully autonomous agents will have the ability to figure out how to hire another agent to integrate the API and then execute the task. From the user perspective, there will be no need to check if an agent can fulfill a task because the agent can determine that themselves.

Bitcoin and AI Agents

In July 2023, Lightning Labs rolled out a proof of concept implementation for using Agents on the Lightning Network called the LangChain Bitcoin Suite. The product is especially interesting as it aims to tackle a growing problem in the web 2 world – gated and expensive API keys for web applications.

LangChain solves this by providing developers with a set of tooling enabling agents to buy, sell, and hold Bitcoin, as well as query API keys and send micro-payments. Whereas on traditional payment rails small micro-payments are cost-prohibitive due to fees, on Lightning Network agents can send unlimited micro payments daily

with minimal fees. When coupled with LangChain's L402 payment metered API framework this could enable companies to adjust access fees to their API as usage increases and decreases, rather than setting a single cost-prohibitive standard.

In a future where on-chain activity is dominated by agents interacting with agents, something like this will be necessary to ensure the agents can interact with each other in a way that is not cost-prohibitive. This is an early example of how using agents on permissionless and cost-efficient payment rails opens the possibilities for new marketplaces and economic interactions.

Outlook

The Agents space is still nascent. Projects are just beginning to roll out functioning agents that can handle simple tasks using their infrastructure – which is often only accessible to sophisticated developers and users. Over time, however, one of the biggest impacts AI agents will have on crypto is UX improvements across all verticals. Transacting will begin to move from point and click to text based, with users having the ability to interact with on-chain agents through LLMs. Already teams like Dawn Wallet are introducing chat-bot wallets for users to interact on chain.

Additionally, it is unclear how agents could operate in web 2 where financial rails are dependent on regulated banking institutions that do not operate 24/7 and cannot conduct seamless cross-border transactions. As Lyn Alden has highlighted, crypto rails are especially attractive compared to credit cards due to a lack of chargebacks and the ability to process microtransactions. If agents become a more common means of transacting, however, it is likely that existing payment providers and applications move quickly to implement the infrastructure required for them to operate on existing financial rails, mitigating some of the benefits of using crypto.

For now, agents are likely to be confined to deterministic crypto-to-crypto transactions where a given output is guaranteed for a given input. Both models, which dictate the capacity of these agents to figure out how to execute complex tasks, and tooling, which expands scope of what they can accomplish, require further development. For crypto agents to become useful outside of novel on-chain crypto use cases will require broader integration and acceptance of crypto as a form of payment as well as regulatory clarity. As these components develop, however, agents are primed to become one of the largest consumers of decentralized compute and zkML solutions discussed above, acting in an autonomous non-deterministic manner to receive and solve any task.

Conclusion

AI introduces to crypto the very same innovations we already see playing out in web2, enhancing everything from infrastructure development to user experience and accessibility. However, projects are still early in their evolution and near-term crypto and AI integration will be primarily dominated by offchain integrations.

Products like Copilot will "10x" developer efficiency, with layer 1s and DeFi applications already rolling AI-assisted development platforms in partnership with major corporations like Microsoft. Companies like Cub3.ai and Test Machine are developing AI integrations for smart contract auditing and real-time threat monitoring to enhance on-chain security. And LLM chatbots are being trained using on-chain data, protocol documents, and applications to provide users enhanced accessibility and UX.

For more advanced integrations that truly take advantage of crypto's underlying technologies, the challenge remains demonstrating that implementing AI solutions on-chain is both technically possible and economically viable at scale. Developments in decentralized compute, zkML, and AI Agents point toward promising verticals that are laying the groundwork for a future where crypto and AI are deeply interlinked.

Contact Us

galaxy.com

For all inquiries, please email contact@galaxy.com.

Legal Disclosure