Introducing Theta EdgeCloud

The first hybrid cloud computing platform built on a fully distributed edge architecture

Theta Labs, Inc / Nov 2023

1. Introduction

Theta Labs introduced the Theta Edge Network and Elite Edge nodes in the Spring 2021 with the launch of Mainnet 3.0, and with the launch of the Theta Metachain in 2023 became the leading decentralized Web3 video infrastructure. The vision was to build out a fully decentralized global edge network run by community members that would provide distributed computing, storage and delivery capabilities for video encoding, transcoding, storage and peer-to-peer delivery. Elite edge nodes that staked TFUEL for “uptime mining” earned additional TFUEL for

delivering higher performance with better hardware and software specs. With a number of partners launching subchains in 2023, edge nodes are now also compensated with partner TNT-20 tokens for various video and AI tasks. Theta Edge Network with its more than 10,000 monthly active nodes has become a key differentiator compared to other L1 blockchains.

The Theta EdgeCloud is the next generation technology that builds on the Theta Edge Network and Theta Metachain, significantly extending and generalizing its capabilities to become the first hybrid cloud computing platform that is built on a fully distributed edge architecture.

This journey began in April 2021 when Theta filed its patent for “Edge Computing Platform supported by smart contract-enabled blockchain network”[1] which was granted in September 2023[2].

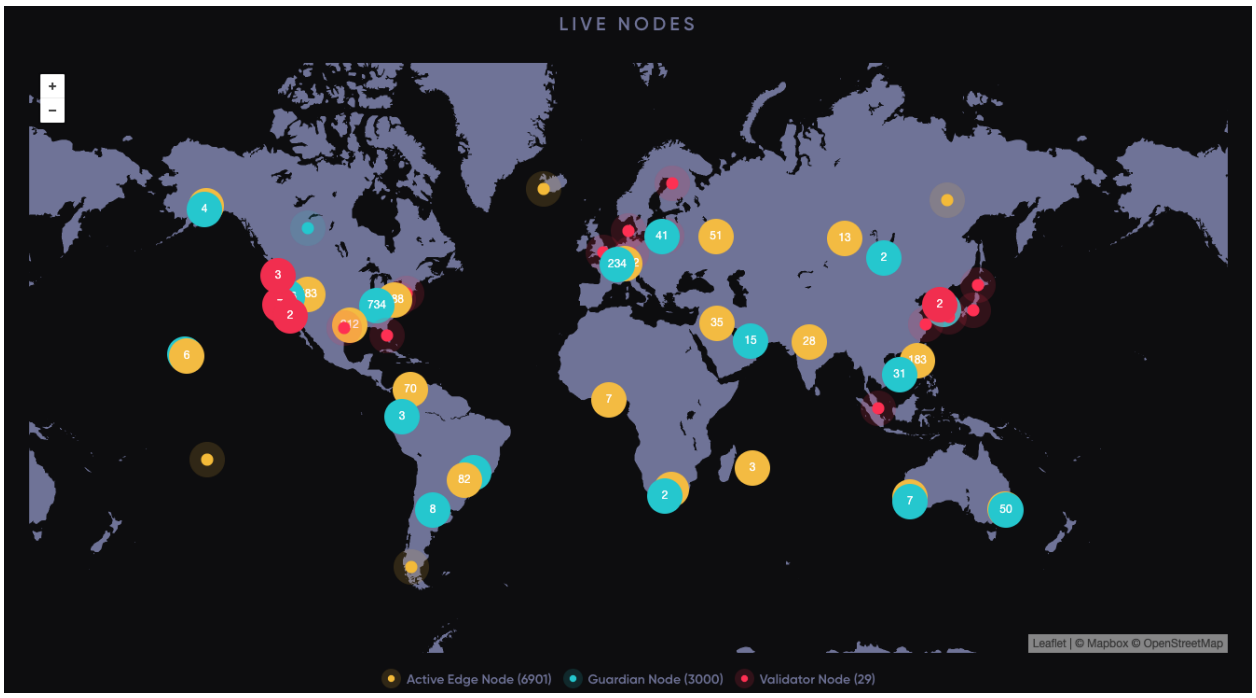

Figure 1. A global distributed network formed by Theta edge nodes serves as the backbone infrastructure for the Theta EdgeCloud.

[1] https://medium.com/theta-network/theta-awarded-us-patent-11-763-332-for-edge-computing-platform-supported-by-smart-contract-enabled-80a994d1248e

[2] https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/11763332

2. Market Opportunity

Theta is introducing the first hybrid cloud edge computing platform that can leverage thousands of edge CPU and GPU nodes worldwide in addition to a hosted, cloud-based distributed architecture. This opportunity is motivated by three major industry trends.

2.1 Increasing video demand globally

The amount of video content and related data continues to grow due to the increasing popularity of social media platforms and rising Internet connectivity particularly in less developed regions. Increasing video data traffic globally is also driven by demand for higher resolution 4K, 8K and higher resolution videos. According to Fortune Business Insights[3] the global video streaming market size was valued at USD $455.5 billion in 2022 and is projected to grow more than four-fold to over USD $1.9 trillion by 2030. This trend drives an increasing need for video encoding, transcoding and delivery services, as well as global visualization and 3D rendering services for various industries including manufacturing, healthcare, construction, education and real estate sectors.

2.2 Adoption of AI across industries and applications

The market for artificial intelligence is expected to grow over 10x in the next decade, driven by adoption of AI applications across verticals from travel to manufacturing. Everything from supply chains, marketing, product making, research, analysis and more are fields that will adopt some aspect of AI within their business structures. The release of ChatGPT in the generative AI market in 2022 was a significant catalyst for end-user adoption of AI and chatbots. The overall AI market is estimated to grow from USD $142.3 billion in 2022 to over USD $1.85 trillion by 2030 according to Statista[4]. This trend drives the need for CPU and particularly GPU compute resources to prototype, train, fine tune and serve AI models.

2.3 Infrastructure compute services

New “bite-sized” infrastructure compute services to support video, media and general AI computation tasks have become increasingly more important as computation data grows exponentially by 10-100x. For example, to fine-tune and train domain specific large language models (LLM) and generative text-to-image and particularly text-to-video AI now requires large amounts of data which makes today’s centralized systems cost prohibitive.

The ability to “paginate” or distribute the workload of vector databases, for example, that store data in a multi-dimensional space with each dimension corresponding to a specific feature or

[3] https://www.fortunebusinessinsights.com/video-streaming-market-103057
[4] https://www.statista.com/statistics/1365145/artificial-intelligence-market-size

attribute, is ground breaking. Vector databases are ideal for similarity search applications like natural language processing, computer vision and recommendation systems. They complement generative AI models well, and provide an external knowledge base for generative AI chatbots to ensure information accuracy.

3. Theta EdgeCloud Overview

3.1 Product Vision

Theta's product vision for 2024-2025 is to deliver a fully distributed hybrid cloud-edge computing platform that caters to the diverse needs of a wide range of users. This platform will be designed to meet the demands of Global 1000 corporations seeking scalability and performance and the flexibility for high growth small and medium businesses (SMBs), researchers and developers at 3-5x cost savings. For example, for a network of high-end NVIDIA 4090s on a per GPU-hour cost basis, Theta's hybrid cloud-edge computing network should be ⅕ to ⅓ of the cost compared to Microsoft Azure, AWS or Google Cloud.

Large enterprises, SMBs, and individuals have distinctive needs. Large enterprises typically require reliable service with service level agreements (SLA) such as guaranteed uptime of at least 99.999%. These organizations operate on a massive scale, serving tens or even hundreds of millions of users, and any downtime or performance issues can have significant business consequences. Thus, reliability, robustness, performance, and privacy are the most important considerations and typically overshadow concerns about cost.

In contrast, SMBs, researchers and community-driven projects often strive for a more delicate balance between cost and service levels. They are often constrained by budget limitations, making them inherently cost-sensitive. Reliable services are important, but they may be willing to make calculated trade-offs between price and service levels to maximize their resources. This flexibility allows them to allocate resources efficiently and prioritize aspects that align with their strategic goals. Theta EdgeCloud's flexibility addresses the needs of any type of user, customizing its usage to serve large enterprises, SMBs, and community projects alike.

3.2 Supported Computing Tasks

The EdgeCloud platform is designed to **support generic containerized compute tasks**, as it implements a virtualization layer where users can run any container on the network. The strength of the virtualization layer lies in its capacity to cater to a myriad of computing needs, all with a unified interface that bridges differing requirements, delivering a robust, scalable, and adaptable infrastructure. Any computation task can be run on this platform as long as it can be containerized, whether it involves complex data analytics, artificial intelligence, video processing, or even specialized scientific simulations. This flexibility offers a dynamic environment where users can harness the power of containerization and distributed edge and

cloud computing resources to optimize, deploy, and scale their workloads with ease and efficiency.

Computing tasks that can be handled by the EdgeCloud platform include but not limited to the following:

- **Media & Entertainment**: As already implemented in the ThetaVideoAPI platform, a decentralized edge network platform can handle resource-intensive media and entertainment tasks such as video encoding, livestream transcoding, and bandwidth sharing efficiently. The Theta EdgeCloud platform will take this capability to the next level. Whether it's encoding video content for multiple formats, transcoding to optimize streaming quality, or rendering complex 3D visual effects for movies and animations, the EdgeCloud platform will provide the scalability and reliability required by the largest media companies.

- **Artificial Intelligence**: AI workflow is fairly complex and involves many computational steps that are ideally suited for EdgeCloud's distributed platform. A typical workflow starts with using Jupyter notebook for data exploration, where data scientists analyze and preprocess the dataset. The next step involves hyperparameter tuning and neural architecture search, where models are optimized for better performance. Subsequent steps include model training, where the chosen deep learning model learns from the data. If needed, fine-tuning and alignment follow, involving further adjustments and iterations to improve the model's accuracy. Finally, when the model is ready for real-world use, it's deployed for serving or inference.

  AI inference applications include generative text-to-image, text-to-video, large language models (LLM), and other custom models. This involves setting up an infrastructure to make predictions accessible to end-users or other systems, often through APIs or endpoints. All these tasks, from Jupyter notebook to model serving, can be easily containerized. In particular, hyperparameter tuning requires the exploration of multiple combinations and configurations, a computationally intensive process that benefits significantly from parallelism. These types of tasks are perfectly suited to run on a highly distributed edge network, where the ability to distribute these computations across multiple nodes or devices can greatly expedite the optimization process and reduce the time required to find the best model hyperparameters.

- **Scientific Research**: Complex scientific simulation jobs include, for example, weather forecasting, climate modeling, astrophysics simulations, chemical reactions, molecular structures, and material properties research. These simulations can also analyze large genomic datasets, conduct DNA sequencing, and perform complex bioinformatics tasks for genetics research in the biomedical field. They require massive computational resources to model physical, chemical, and biological phenomena accurately.Using a hybrid of edge and cloud resources, Theta EdgeCloud can direct simulation jobs to the best combination of compute resources to efficiently complete the task.

- **Financial Modeling**: In the financial sector, complex models are often used for risk analysis, high-frequency trading, and Monte Carlo simulations to assess investment strategies and market behaviors. These applications play a pivotal role in the industry to analyze and optimize investment strategies, market trends, and risk management. The computational demands of these financial models can be substantial, requiring access to high-performance computing resources capable of handling large-scale data and intricate calculations in real-time. The EdgeCloud platform's distributed highly parallel architecture is ideally suited for addressing these types of financial computational needs.

4. EdgeCloud Architecture

To realize the product vision discussed above requires a flexible and scalable hybrid cloud edge computing architecture as shown in Figure 2.

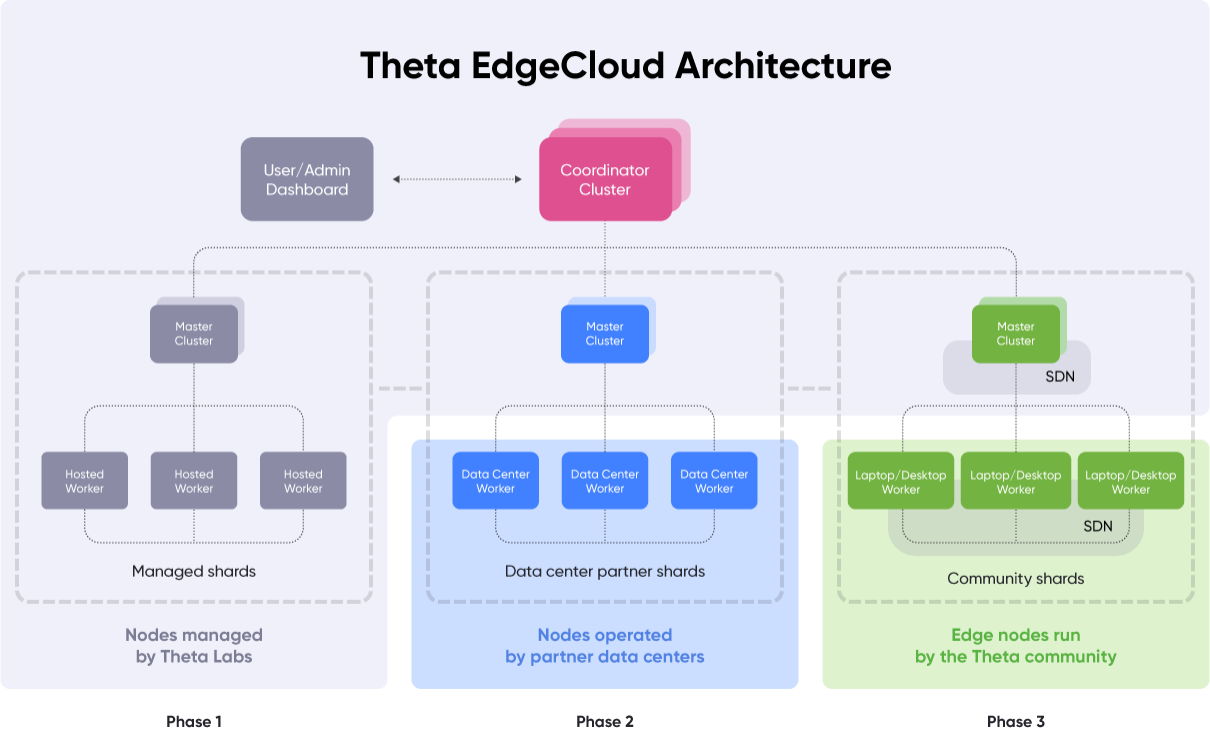

Figure 2. Theta EdgeCloud Architecture and a phased release strategy.

4.1 Multi-shard Architecture

The key design principle is to divide the edge nodes into multiple shards, each type catering to different business requirements. For example, enterprise computation tasks need to run in a private and secure environment, versus non-sensitive NFTs, media files, and generative AI tasks can be executed in a public environment. Other considerations include node reliability, robustness, and performance. Enterprise customers also have more strict SLAs so those tasks are better run on more reliable nodes. Figure 2 outlines three types of shards technically:

- **Managed Shards**: The workers of this type of shard are nodes hosted by Theta Labs in the cloud environment. The worker nodes can communicate directly with each other through a private network. A managed shard is more reliable and ideal for workloads like AI model training, jupyter notebook and tasks with strict SLA requirements from enterprise customers.

- **Data Center Partner Shards**: These data center workers are nodes run in partner data centers. The worker nodes can communicate directly with each other, and can communicate with the master cluster through public IPs. Data center shards run by crypto mining businesses, regional cloud service providers and other partners are also more reliable and equipped with better hardware and software, suitable for workloads with certain SLA requirements.

- **Community Shards**: The workers of this type of shard are community nodes running on laptops and desktops which typically do not have a public IP address. For the worker nodes to communicate directly with each other, a software-defined network (SDN) network can be configured. This type of shard is typically less performant, reliable and robust. However, they are optimal for smaller, faster, and highly parallelizable jobs (e.g. AI hyperparameter tuning), and tasks like AI inference and video encoding with built-in fault-tolerant mechanisms. In the long-run, community nodes would also encompass a wide range of devices including smartphones, smart TVs and distributed IoT devices.

4.2 Shard Management

To effectively coordinate the shards and the worker nodes, a hierarchical worker node management approach will be implemented:

- A **“coordinator cluster”** oversees and manages the entire infrastructure, allowing for infinite scalability and fault tolerance. It is in charge of intelligently assigning and routing a compute task to the most optimal cluster-node.

- Within each shard, a **“master cluster”** manages and orchestrates that shard instance consisting of an orchestration control plane providing an API server, scheduler and other control components. The master cluster will also implement multiple master nodes to

ensure fault tolerance.

- A **“worker node”** actually runs the Theta edge node client software. This will be consistent across all shards, except for its hardware specifications and performance. For example, in community shards, worker nodes are composed of lower and mid-tier CPUs and GPUs on laptops, PCs, smartphones and other IoT devices. By contrast, in managed shards initially operated by Theta, worker nodes will incorporate high-end NVIDIA GPUs such as Tesla A100s/H100s with up to 312/624 FP16 TFLOPS, T4s with 65 FP16 TFLOPS, and others.

4.3 Task Assignment

Compute tasks are initiated by EdgeCloud customers and partners through a user-facing dashboard or API for developers. The task metadata is passed to the coordinator cluster, which then decides which shard and worker nodes the task should be assigned to. In order to intelligently optimize the distribution of tasks, an efficiency score will be assigned to each worker node across the EdgeCloud network, a function of various factors:

efficiency_score = *f(job_type, price, latency, availability, computing capacity, fairness)*

Given a particular computing task, the coordinator will compute in real-time and route the task to the node and shard with the highest **efficiency_score**. More advanced technologies like reinforcement learning could also be adopted to determine the task assignments and improve overall system efficiency. We have conducted preliminary research on this technology and published our research findings at the prestigious ACM SIGKDD 2023 conference[5].

[5] Xing et al., A Dual-Agent Scheduler for Distributed Deep Learning Jobs on Public Cloud via Reinforcement Learning, ACM SIGKDD 2023

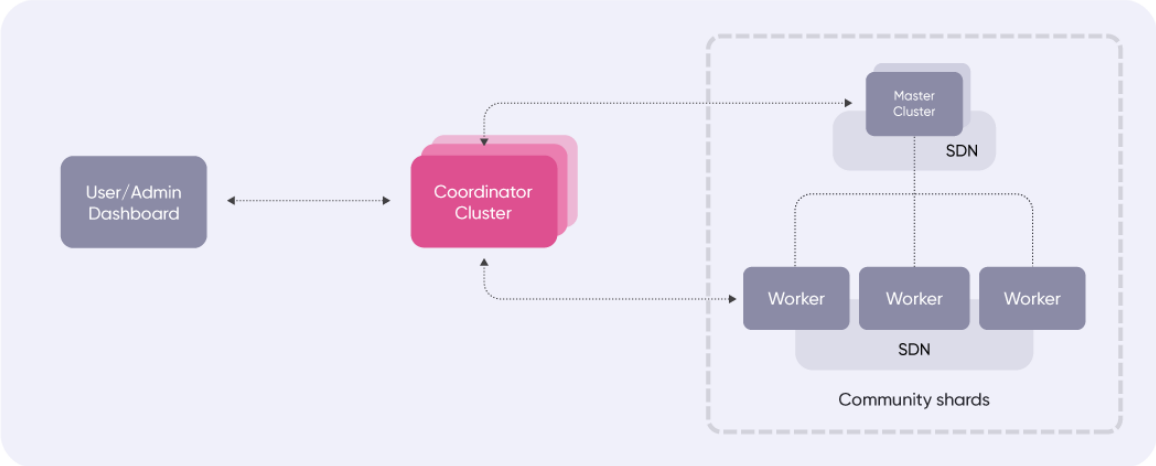

Figure 3. The task assignment process. A customer initiates the compute task through the user-facing dashboard or API, which is then assigned by the coordinator cluster to a worker node and shard with the highest efficiency_score.

4.4 Token Rewards for Edge Nodes

Edge nodes are rewarded for compute jobs they perform with either TFUEL or TNT-20 tokens issued by Theta partner projects built on the Theta Metachain. The coordinator will implement a reward module to calculate job rewards based on various factors, including but not limited to 1) job type, 2) correctness of the computation results, 3) job completion time, and 4) resource usage of the edge node.

To prevent malicious edge node operators from gaming the system for token rewards, the coordinator will check the correctness of the computation results. Verification of correctness can be accomplished through the implementation of any of the following approaches:

- **Zero-knowledge based correctness checks**. Zero-knowledge technology allows the public to check whether a task is completed as prescribed without revealing any specific details about the task itself and without disclosing the actual data or process involved. This is valuable for privacy and security, as it ensures that sensitive information remains confidential while still confirming the accuracy of the work performed.

  With recent advancements in zkSNARK and zkML, these proofs can be made very short and fast to verify. As these technologies mature, a worker client can potentially leverage them to generate succinct correctness proofs for the computation jobs assigned, and upload the proofs along with the results to the coordinator. The coordinator then verifies the proof and rewards the worker only when the proof is valid. On this front, Theta Labs

collaborated with FedML earlier on a related research project, which resulted in a research paper published at the IEEE DAPPS 2023 conference[6].

- **Cross-check**: Given the vast parallel computing power of the EdgeCloud platform, it is possible to assign the same jobs to multiple nodes, and cross-check the results.
- **Server side random verification**: The coordinator can also randomly sample a certain percentage of jobs and rerun those jobs on trusted machines. By comparing the result, the coordinator can evaluate whether the result uploaded by a worker node is valid or not.

5. Additional Considerations

5.1 Support for Third Party Client Integrations

As an open network, the Theta EdgeCloud platform will support third party compute client integrations in addition to first party workers. Currently, the edge node software has incorporated a few third party AI clients including FedML and Lavita, in addition to native video encoding, transcoding and delivery clients.

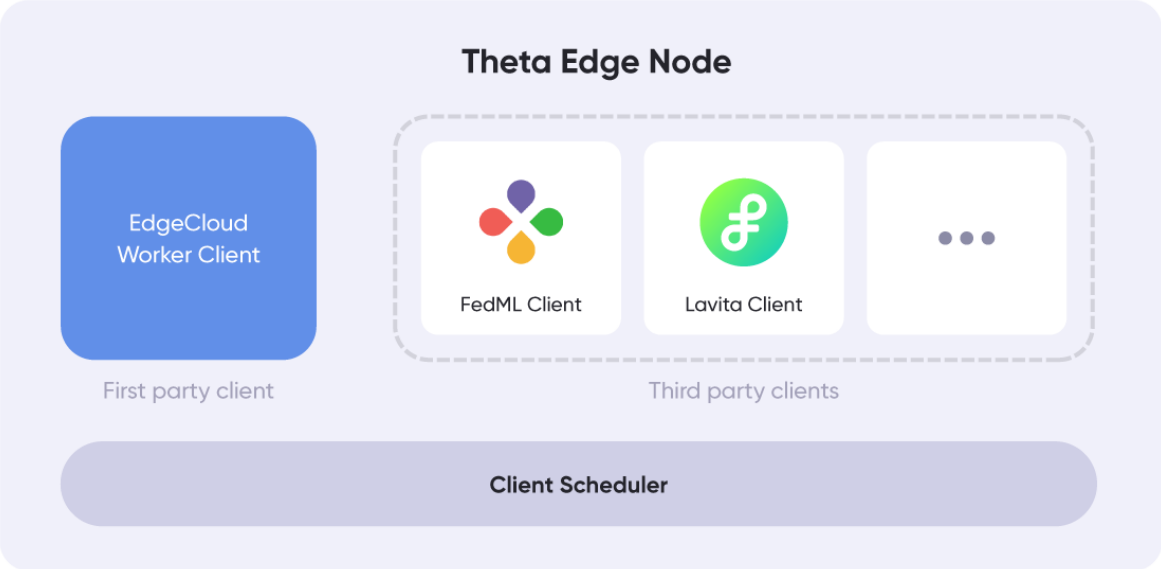

Figure 4. Theta Edge Node supports both first party and third party clients. The Edge Node software also implements a client scheduler which launches and coordinates various clients.

[6] Buyukates et al., "Proof-of-Contribution-Based Design for Collaborative Machine Learning on Blockchain", IEEE DAPPS 2023

FedML is a cloud service for LLMs & generative AI. With the FedML client integration, users can tap into the Theta EdgeCloud through the FedML.ai portal and launch complex model training, deployment, and federated learning on decentralized GPUs, multi-clouds, edge servers, and mobile devices. The FedML client and reward system integration with Theta edge nodes was completed in November 2023. The edge nodes are currently compensated with TFuel rewards for the FedML AI jobs.

Lavita is a blockchain-AI based next generation healthcare platform, enabling the use of massive biomedical datasets for research while preserving individual privacy and ownership of data. Through its integration with Theta EdgeCloud, the Lavita platform connects individuals with data miners such as pharmaceutical and research companies to participate in new drug trials and treatments, gain insightful genetic information and latest research findings based on their clinical needs. Launched in August 2023, Theta edge nodes earn Lavita Token rewards for healthcare AI model training.

Over time many more third party clients will be integrated into the Theta EdgeCloud platform and serve as important building blocks to improve system robustness, fault tolerance, and offer alternative ways for users to access the EdgeCloud computational power. As these clients continue to evolve and add more features, they will further enhance the overall functionality and utility of Theta EdgeCloud.

5.2 Distributed Storage Roadmap Implications

Distributed storage is an integral part of the new Theta EdgeCloud platform. While Theta EdgeStore alpha release currently offers storage service in the form of a KV-store, in the EdgeCloud context, the distributed storage service could take different forms. Below are different types of use cases for distributed storage in a compute-intensive scenario:

- **Temporal storage**: This use case requires retaining data in memory or disk for short durations, for example the training data required for AI models. This enables quick access to critical information, enhancing the efficiency of compute-intensive jobs including machine learning and data analysis.

- **Persistent storage**: The EdgeCloud storage system can be used to host distributed databases, which requires persistent as well as updatable storage. One such type of database that has gained significant traction among the machine learning community in recent years is the so-called “vector database”. It is ideal for applications that require fast and accurate similarity search, such as recommendation systems, natural language processing, and computer vision. Vector databases make it easier for machine learning models to retrieve external knowledge based on similarity searches.

- **Immutable storage**: Traditional media files including video VoD, and NFT data files require immutable storage. Immutability means that once a file or digital content is

stored, it cannot be altered, tampered with, or deleted. This is crucial for NFTs and media files because they represent ownership or provenance of unique digital items or content such as digital art, collectibles, or other virtual assets.

In many cases a worker node will need to access local storage while performing a compute task. In the use cases listed above, a machine learning algorithm needs to access and load the training data while in the process of training a model. During model inference, it might need to query the vector database for external knowledge. In the case of video transcoding, the original source video file needs to be loaded in order to run transcoding algorithms. Our system design will account for storage resources utilized in a temporal, persistent or immutable mode, where compute rewards will include compensation for the associated storage utilized.

6. Product Launch Strategy

Theta EdgeCloud will roll out in three releases to address immediate market opportunities and minimize technical risks while implementing the architecture in Figure 2:

- Release 1: Launch **cloud hosted nodes managed by Theta** integrating community token incentives and rewards. This enables Theta to test the fully distributed architecture in a controlled environment and allows for high levels of SLAs for initial enterprise customers.

- Release 2: Enable **partners including cloud service providers, mining businesses and community members to set up and run cloud hosted nodes in partner data centers**. This allows traditional cloud businesses to integrate Theta EdgeCloud into their existing product offering, providing more cost-effective options to their customers.

- Release 3: Roll out **fully distributed architecture to Theta edge nodes run by the community working seamlessly with release 1 and 2 networks**. A “coordinator cluster” routes any incoming computing task to the most optimal node based on type of job, immediacy, performance, cost and other service levels parameters.

6.1 Release 1 Overview - Cloud Hosted Nodes by Theta

Release 1[7] will enable elite edge node (EEN) operators to participate in EdgeCloud nodes hosted by Theta Labs via a “lock and earn” mechanism. A new **“Elite+ Booster”** section will be unlocked when that edge node stakes the full 500,000 TFUEL through a single non-delegated wallet. This Elite+ node can then lock additional TFUEL (separate from the 500,000 for elite status) and earn boosted TFUEL rewards for all compute tasks performed in the EdgeCloud. There is no limit to the additional amount of TFUEL that can be locked via Elite+ Booster option and bonus reward multipliers will apply to longer time locks.

[7] Currently scheduled for summer 2024 launch (subject to change)

EdgeCloud hosted nodes by Theta will initially include high-end NVIDIA GPUs such as Tesla A100s, H100s with up to 312/624 FP16 TFLOPS, T4s with 65 FP16 TFLOPS, and other high performance GPUs. These are ideally suited for industrial scale video, large language model AI, generative AI and other GPU-intensive computation tasks for the scientific and financial communities, for example.

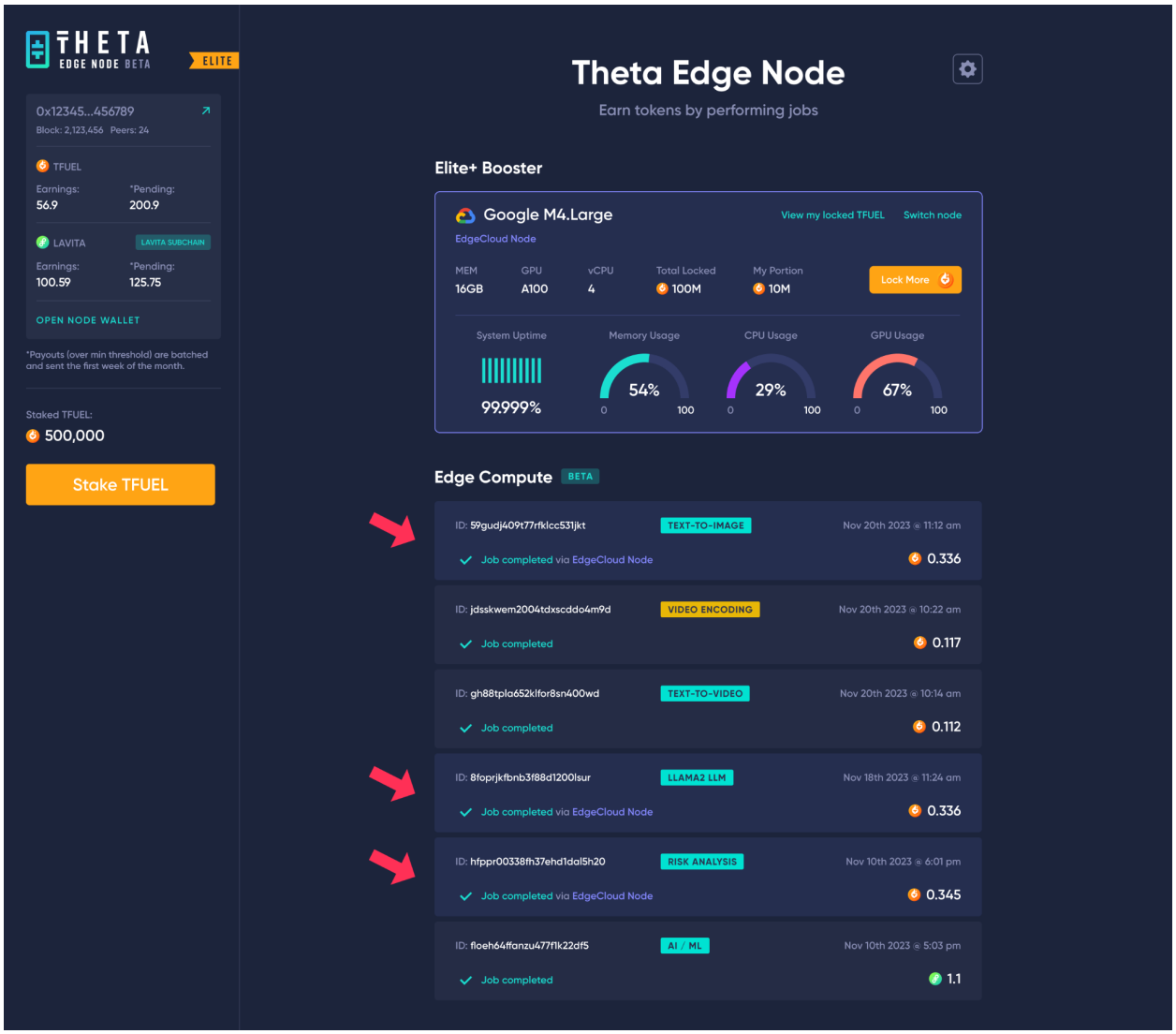

Figure 5. Mockup of Elite edge node client software with “Elite+ Booster” option for Release 1[8]

Each EEN can select their preferred EdgeCloud node to support, and can switch to another node at any time based on their performance, earnings and other reasons. Users will have access to an operations dashboard which is effectively a virtual management window into the

[8] For illustrative purposes only, design and layout subject to change

node to view jobs completed/type/payouts, performance and more. A particular EdgeCloud node can have N number of EEN supporting it, and all earnings will be shared proportionally among all TFUEL locked by the N elite edge nodes.

In the long-term, any EEN supporting a particular EdgeCloud node will have the option to collaborate and vote on various upgrades and enhancements to the underlying cloud node hardware and software specifications, e.g. increasing VRAM from 8GB to 16GB, etc.

7. Conclusion

Theta's vision for a decentralized video compute, storage and delivery edge architecture was introduced in 2021 with the launch of the Theta Edge Network powered by TFUEL and further extended with the launch of Theta Metachain in 2023. That vision is now coming to fruition with over 10,000 monthly active edge nodes hosted by community members around the world earning TFUEL and TNT-20 rewards and over 2.4 Billion[9] TFUEL staked against Elite Edge nodes. Building on this solid foundation, the introduction of Theta's next-generation generalized EdgeCloud architecture in 2024-2025 could fundamentally transform today's centralized cloud service businesses.

The immediate strategy is to partner and align with some of the largest cloud providers and become part of their extended architecture making Theta indispensable to their business.

- This begins with Release 1 of Cloud hosted nodes managed by Theta and extending the edge node client with an "Elite+ booster" option. With this first step, all elite node operators will be able to participate in the initial development and deployment of Theta EdgeCloud's distributed architecture, supporting high performance CPUs and GPUs.

- Release 2 extends this capability for any data center, mining operator or community member to deploy these in their own cloud vastly increasing the capability, robustness and fault tolerance of Theta EdgeCloud.

- Finally, release 3 integrates and introduces the world's first hybrid cloud computing platform built on a fully distributed edge architecture.

With today's accelerating adoption of decentralized Web3 video, NFT gated video content with digital rights management[10], generative AI and other GPU-intensive applications, this is the perfect time to launch Theta EdgeCloud and define the future of decentralized computing.

[9] https://explorer.thetatoken.org/ as of Nov 19, 2023

[10] https://medium.com/theta-network/theta-awarded-patent-application-approval-for-uspto-17-218-245-theta-1003-decentralized-drm-9f41c4459b83 NFT DRM digital rights managent patent granted in 2021