Infera Network Litepaper

Infera Team

April 2024

Table of Contents

1 Introduction

The field of artificial intelligence (AI) is currently experiencing a significant technological transformation that is marked by a rapidly increasing demand for computational power in the developer space and the consumer ecosystem. Despite the abundant availability of consumer-grade GPUs, a significant portion remains underutilized in the space. At the same time, advancements in AI models have enabled custom deployments on consumer hardware, mitigating the dependency on high-cost data center GPUs. Addressing these developments, Infera Network is pioneering a decentralized platform that democratizes participation in the AI economy. The platform connects readily available GPUs by narrowing the accessibility gap for developers through incentives based participation on the Infera Network. Infera aims to empower the AI economy by providing a development experience that matches the efficiency of centralized solutions while harnessing the advantages of decentralized computing infrastructure.

2 Problems

The heightened demand for data center GPUs has created extensive waitlists for developers and researchers alike, presenting critical challenges in accessing vital computing resources. Such delays are a stark indicator of the need for a more effective and robust allocation and distribution of GPU resources to power these new AI applications. Despite various attempts to alleviate the issue, GPU scarcity remains high and signifies an acute need for better compute distribution to consumers.

Another avenue to explore is the pronounced cost differential between data center GPUs and their consumer-grade counterparts. Data center GPUs are generally priced 10x higher than consumer models imposing significant financial hurdles for hobbyists or startups seeking GPU computation. This disparity highlights an avenue for a new alternative approach, including the mobilization of underutilized consumer hardware, to narrow the accessibility divide and availability of computational resources across a more extensive user base.

3 Unique Value Proposition

Infera provides developers with an efficient pathway to AI inference services and opens up new avenues for revenue generation for node operators. Within the dynamic context of GPU deployment, where leading cloud providers have transitioned from cryptocurrency mining towards offering AI computational services, Infera levels the playing field. This transformation affords individual GPU holders the opportunity to engage in the AI market, challenging the monopolistic hold of major entities in the sector. By adopting a token-based remuneration system, Infera departs from the traditional pay-per-hour GPU usage model. It introduces a mechanism where developers can utilize tokens to access computational services, and node operators are compensated for their contributions of processing power to the network. This model promotes not only enhanced cost-effectiveness and flexibility in pricing but also encourages wider participation. Consequently, it cultivates a more inclusive ecosystem, where the value generated from advancing AI technologies is equitably shared among all stakeholders.

4 Use Case and Application

Infera is designed to address the growing demand for accessible and efficient AI solutions, particularly focusing on the utilization of Large Language Models (LLMs). Inferas approach to decentralizing AI capabilities leverages the power of distributed computing to offer scalable and cost-effective processing.

In the short term, the Infera Network is poised to revolutionize various sectors by facilitating access to Large Language Model (LLM) inference capabilities. We hope to see a plethora of new applications of this exciting new tech, from enhancing user interactions for businesses to the refinement of data analysis.

Chatbots leveraging LLMs are an obvious use case that can be programmed to interact and assist humans. Chatbots are built with the purpose of servicing a wide range of roles. From engaging as a customer service representative and engaging in complex conversations, to providing users with a cooking recipe, chatbots can improve the overall quality of life of humans. Another interesting use case is using text embeddings. Text embeddings are critical for vector search operations and database functionalities. This feature allows for the efficient comparison and retrieval of text data, enhancing the effectiveness of search engines and recommendation systems.

Language translations and text summarization both utilize LLM inference. With LLM inference, it is possible to facilitate seamless language translation, breaking down linguistic barriers and introducing a new way for people to interact with each other, cross culturally. Additionally, text summarization allows for users to digest large pieces of information through a summary generated by the LLM. AI inference makes it possible to summarize extensive documents into concise and information rich summaries, without losing details and context. Implementation into group chats for creating meeting minutes or to find free time for meetings through contextualization, the possibilities are endless.

For assisting developers, apps built on top of Infera can help enhance developer productivity through assisting with code generation from natural language descriptions. This would help support rapid prototyping and development cycles for software development. Infera can also assist in producing diverse content across various domains, such as writing skits to creating cooking recipes. Aiding content creators and marketers in generating innovative and engaging material is an avenue the Infera incubator hopes to support.

Looking ahead in the long term, Infera aims to facilitate new models in a decentralized manner, incorporating advanced techniques as they mature. This vision extends to various generative AI applications, including generative AI and speech to text as the network continues to mature. Exploring into the realms of image, speech, and video generation, as models and techniques evolve, the network will enable these capabilities to be more widely accessible. As models for speech to text and video generation mature, they will be integrated into the network, further expanding the scope of AI applications that Infera has to offer.

Fine-tuning pre-trained LLMs for specific tasks or datasets can significantly enhance their performance. The Infera Network offers opportunities for developers to monetize their expertise in fine-tuning models across various domains such as sentiment analysis, text classification, named entity recognition (NER) and dialog systems.

Sentiment Analysis: By analyzing texts from social media, customer reviews, and news articles, fine tuned LLMs can provide unique insights from targeted and tailored data sets. Paired with other data sets, custom signals can be generated for trend analysis.

Text Classification: Useful for categorizing documents and emails for users and routing support tickets which can enhance productivity if built together with an integrated calendar system.

Named Entity Recognition (NER): Essential for extracting and identifying specific entities from texts, aiding in information retrieval and analysis in sectors like healthcare and finance.

Dialogue Systems: Development of domain-specific conversational agents that provide personalized interactions, enhancing user experiences.

Infera is excited to distribute and democratize the access to these AI technologies through our platform. Along with a vision for future generative AI applications, Infera is setting the stage for a new era of innovation and accessibility in AI. By providing a platform for LLM inference and fine-tuning, Infera hopes to build a network of AI applications ecosystem that's used by all.

5 Infera API & Network Fees

The Infera API stands at the forefront of the network's developer-focused offerings, embodying the principle of seamless integration and broad compatibility with existing AI frameworks. Crafted to adhere to OpenAI's REST API standards, it ensures a frictionless transition for projects migrating to decentralized AI computations. The use of an auth/bearer token for authentication streamlines access, maintaining a high standard of security and privacy for developers leveraging the network's resources. Additionally, the option for an on-chain oracle verifies outputs directly on the blockchain, presenting an unparalleled level of trust and reliability in the result provided by Infera's API.

Infera Network introduces a transparent fee structure, designed to support the network's sustainability while maximizing value for its stakeholders. Transactions from developers will incur a processing fee, with the majority of the network's fees distributed to support the operations and development of the chain. Specifically a portion of the fees is dedicated to buying and burning $INFER tokens from the public LP. Another segment funds ongoing research and development to help continue to grow the ecosystem. This includes the Infera incubator program.

6 Infera Incubator

The Infera Incubator program represents an initiative designed to foster innovation and support developers with their startups and projects with the Infera network. Recognizing the challenges faced by developers in bootstrapping AI projects, Infera aims to provide resources , guidance and access. The program is tailored to assist developers in navigating the complexities of AI development, offering a springboard to bring their visionary AI project to life on the Infera platform.

Mentorship and Support: Infera plans on opening up its network of AI researchers, developers, and industry veterans that are working with Infera, providing insights into best practices, technical strategies, and market positioning to help projects scale their solutions and accelerate growth. The incubator program is also in place to help connect promising projects with potential investors and partners, facilitating funding opportunities that can help scale their solutions and accelerate the growth of their user base. Developers will become part of a vibrant community of AI enthusiasts, researchers, and entrepreneurs, fostering collaborations and opening doors to valuable network opportunities.

Hackathons: In parallel with the incubator program, Infera plans to host a series of hackathons aimed at highlighting the platform's capabilities while nurturing new and upcoming software teams. These hackathons will serve as a dynamic platform for developers to showcase their creative, technical skills, and innovative solutions leveraging the Infera network's unique offerings.

Visibility and Recognition: With the incubator and hackathon, teams and standout projects will receive recognition and exposure for attending our events and interacting with the community. Infera aims to accelerate the development of AI applications and contribute to the community in a meaningful way. Infera has developed partnerships with companies to showcase tools and companies built on top of Infera, helping Infera ecosystem projects gain the initial traction they need.

7 Conclusion

In summarizing the insights and objectives outlined, Infera represents a pivotal advancement in the convergence of artificial intelligence and blockchain technology. By decentralizing the computational processes that underpin AI through a distributed network of GPUs, Infera is setting a new standard for how these technologies can be leveraged collaboratively, offering a scalable and efficient platform that addresses the current limitations faced by developers and businesses alike.

Central to Infera's mission is the provision of an ecosystem that not only simplifies access to computational resources but also supports the development and deployment of AI applications at scale. The Infera Incubator Program underscores our dedication to this cause, aiming to foster innovation by providing developers with the resources, guidance, and network needed to excel. Concurrently, our hackathons are structured to bring people together to test, build, and refine new ideas in a collaborative environment. The technical foundation of Infera is solidified by our API, designed for high compatibility and security, facilitating straightforward integration for developers accustomed to working with OpenAI, allowing for an easy and seamless transition.