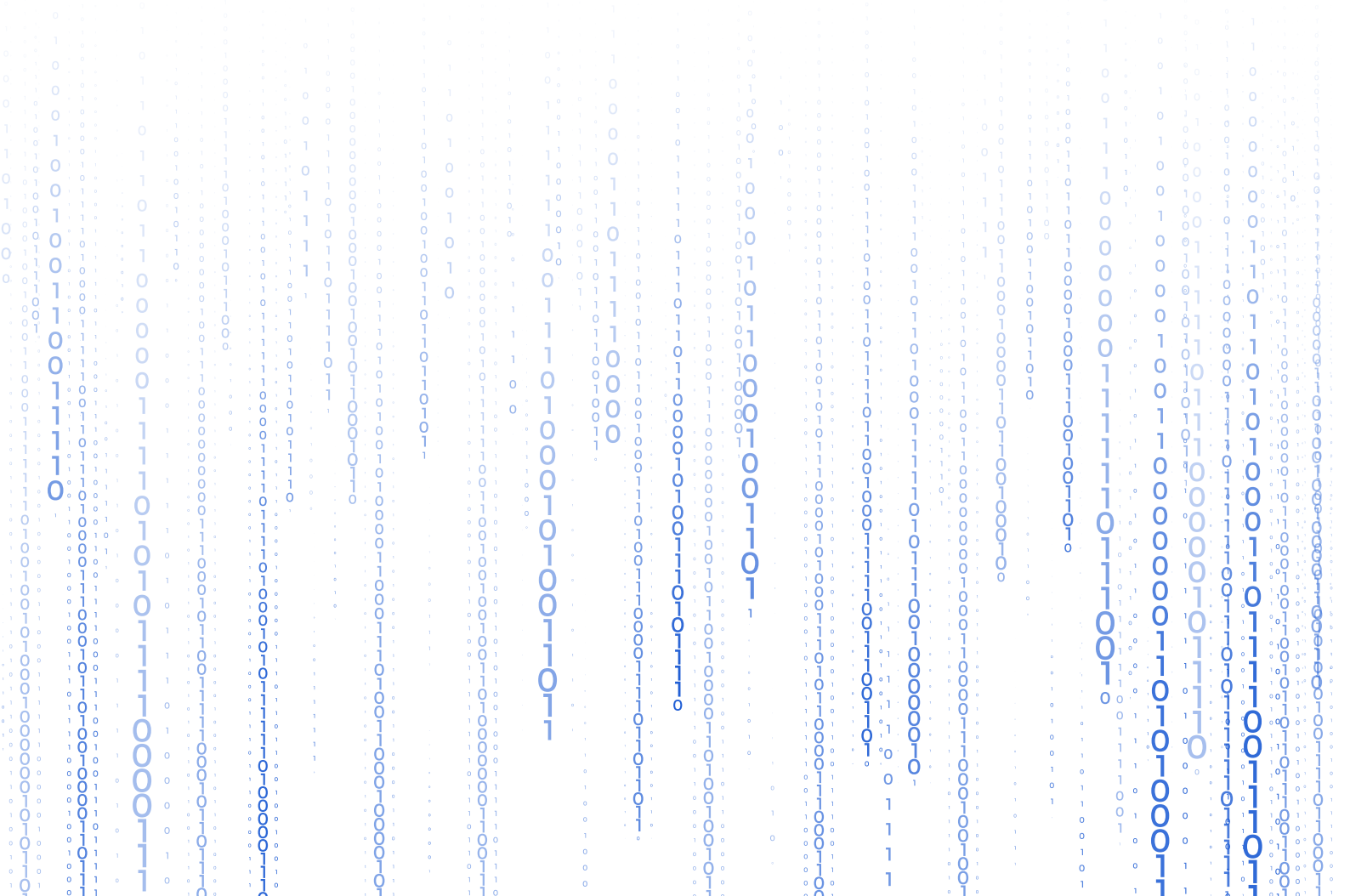
HOC

House of Chimera Research: Latest Developments on GPU Computing

Our research paper aims to offer balanced insights into the latest developments in distributed GPU computing within the Web3 industry, highlighting one of the newest projects in this domain, io.net.

May, 2024



The content is for informational purposes only, and you should not construe any such information or other material as legal, tax, investment, financial, or other advice. Nothing contained in the research paper constitutes a solicitation, recommendation, endorsement, or offer by House of Chimera or any third party service provider to buy or sell any securities or other financial instruments in this or any other jurisdiction in which such solicitation or offer would be unlawful under the securities laws of such jurisdiction. All content of the research paper is information of a general nature and does not address the circumstances of any particular individual or entity. Nothing in the research paper constitutes professional and/or financial advice, nor does any information on the research paper constitute a comprehensive or complete statement of the matters discussed or the law relating thereto. House of Chimera is not a fiduciary by any person's use of or access to the research paper. You alone assume the sole responsibility of evaluating the merits and risks associated with using any information or other content of the research paper before making any decisions based on such information. In exchange for using the research paper, you agree not to hold House of Chimera, its affiliates, or any third-party service provider liable for any possible claim for damages arising from any decision you make based on information or other content made available to you through the research paper.

House of Chimera is an independent blockchain research and advisory firm, committed to integrity and transparency. We are fully transparent about our holdings and personal interests within io.net. House of Chimera has a financial stake in io.net through our investment in the io.net native token, IO. Our integrity remains uncompromised in researching io.net, as the io.net team had no influence on our research at any stage.

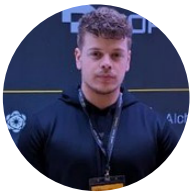
HOC

Contents

HOC

Foreword

This paper explores the dynamic evolution of distributed GPU computing in the Web3 industry, highlighting io.net as a pioneering project. We delve into the increasing demand for high-performance computing resources driven by AI advancements, and the opportunities and challenges faced by decentralized cloud providers in this rapidly growing market.



Diederick Jacobs

Founder

House of Chimera

HOC

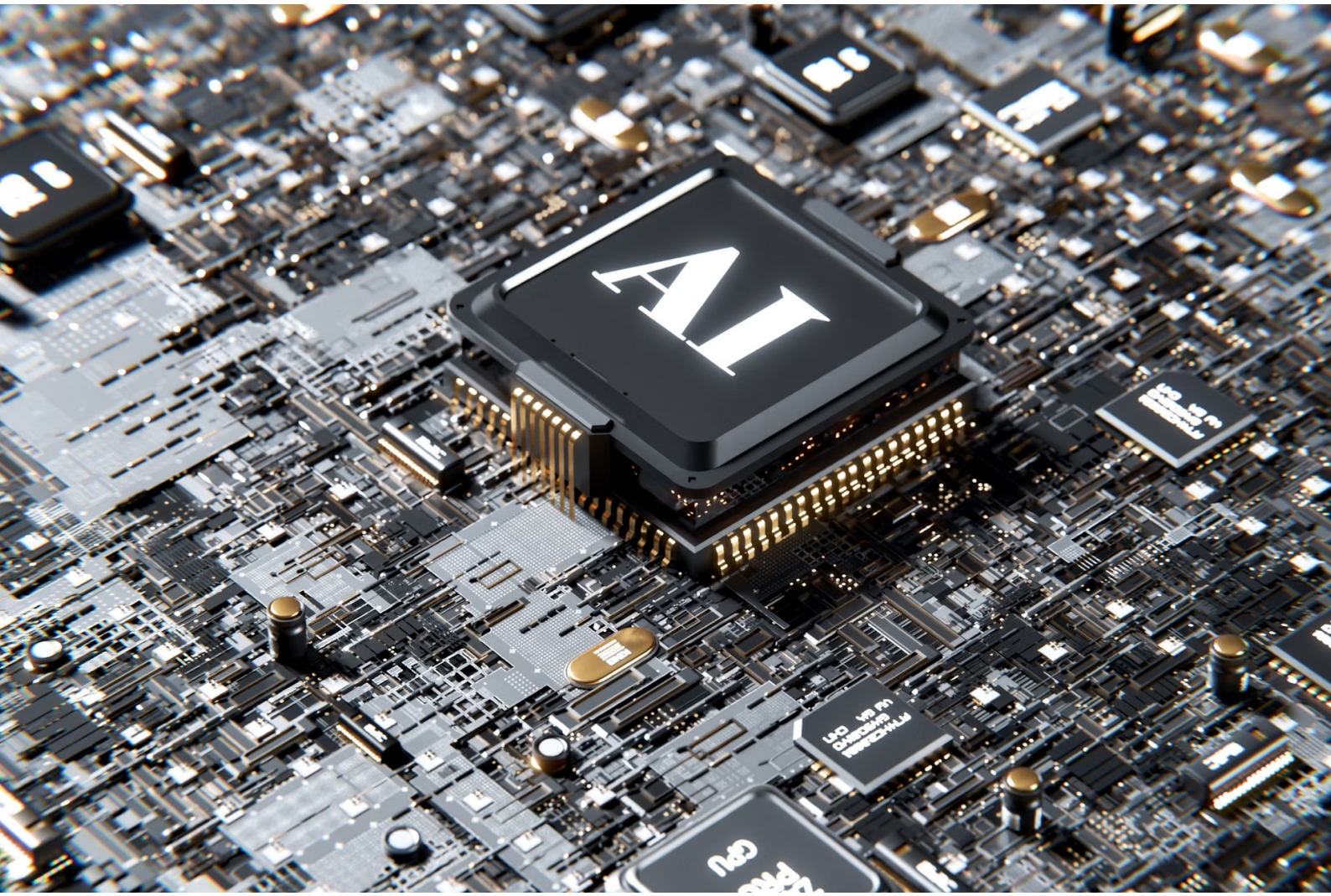


Photo by Igor Omilaev on Unsplash

The Rise of AI and GPU Computing

Exploring the growing demand for high-performance computing resources and the rise of io.net as a decentralized solution in the AI-driven landscape.

In the first half of 2023, OpenAI captured headlines with the rise of ChatGPT, as the Artificial Intelligence (AI) industry, currently valued at approximately $180 billion, is expected to grow to over $800 billion by 2030.[1] ChatGPT and most other AI models are Large Language Models (LLMs), notable for their ability to achieve general-purpose language generation. These models can learn, comprehend, and generate human language text, allowing them to produce text, images, videos, and more by utilizing vast amounts of training data.[2] These advances have formerly opened the door for major tech companies (e.g., Google, Meta, X) to offer LLM-based applications.

To improve the Deep Learning (DL) algorithms of these LLM applications, increasing computing power is essential, further driving the demand for computational resources. LLMs require vast amounts of high-end Graphics Processing Units (GPUs) and Central Processing Units (CPUs) for training and inference. Training a LLM necessitates massive computational resources as the model utilizes billions of parameters, making efficient data structuring and retrieval vital. This rising demand is particularly prominent among tech companies, AI startups, and cloud providers. Cloud providers are acquiring as many GPUs as possible to offer their customers access to these resources for running AI workloads.[3] Big tech companies like Meta and Tesla have significantly increased their purchasing of custom AI models and internal research.[4]

Foundation model companies like Anthropic[5] and data platforms like Snowflake and Databricks have acquired a significant number of GPUs to support AI development and their core operations. To meet this surging demand, major GPU providers have introduced numerous hardware innovations to improve AI-optimized computing. Nvidia has enhanced its GPU microarchitecture to tightly integrate tensor cores, specialized units for matrix operations targeting DL workloads. Additionally, several AI Application-Specific Integrated Circuits (ASICs) have been announced, such as Groq's Tensor Streaming Processors[6] and Graphcore's Intelligence Processing Unit[7], promising even better performance for AI-computing applications.

The sudden increase in demand has led to a significant supply crunch, with delivery times peaking at 11 months. Currently, delivery times are around 4 months, which is still considerable

[1] Statista Market Insights (2024, March). Artificial Intelligence - Worldwide: Market Size. Statista. https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide#market-size.

[2] IBM (n.d.). What are large language models (LLMs)? https://www.ibm.com/topics/large-language-models.

[3] Gardizy, A. (2023, December 7). Why Amazon and Nvidia are Teaming Up in the Cloud? The Information. https://www.theinformation.com/articles/why-amazon-and-nvidia-are-teaming-up-in-the-cloud.

[4] Garreffa, A. (2024, February 14). Tesla will spend billions of dollars on NVIDIA AI GPUs this year, will also buy AMD AI GPUs. TweakTown. https://www.tweaktown.com/news/95840/tesla-will-spend-billions-of-dollars-on-nvidia-ai-gpus-this-year-also-buy-amd/index.html.

[5] Morgan, T.P. (2024, March 27). Amazon gives Anthropic $2.75 billion so it can spend it on AWS XPUS. The Next Platform. https://www.nextplatform.com/2024/03/27/amazon-gives-anthropic-2-75-billion-so-it-can-spend-it-on-aws-gpus/.

[6] Southard, D. (2019). Tensor Streaming Architecture delivers Unmatched Performance for Compute Intensive Workloads. Groq White Paper, 1-7.

[7] Jia, Z., Tillman, B., Maggioni, M., & Scarpazza, D.P. (2019). Dissecting the graphcore ipu architecture via microbenchmarking. arXiv preprint arXiv:1912.03413, 7-91.

due to the sustained high demand from AI and cloud companies, which account for approximately 60% of the demand.[8] This surge in demand prompted Nvidia CEO Jensen Huang to address the allocation of GPUs, amid accusations that Nvidia primarily distributes its high-end GPUs to cloud service providers.[9] This has raised concerns toAI startups and innovators being cut off from these resources, significantly reducing their competitiveness. Additionally, there have been instances where certain countries were banned from accessing GPUs, effectively censoring these countries from high-end GPU resources.[10] The primary reason for banning high-end GPUs in certain countries is to prevent advancements in AI, with China specifically being targeted to restrict access to these advanced technologies.

Io.net offers an enterprise-grade decentralized computing network that enables ML and DL engineers to access distributed cloud clusters at a fraction of the cost of comparable centralized services. This approach grants them access to high-end GPUs and CPUs, which are often out of reach due to high costs, limited availability, or significant overhead. This paper aims to provide context and insight into io.net while also evaluating the industry and its competition.

In the first half of 2023, OpenAI captured headlines with ChatGPT as the AI industry, valued at $180 billion, is projected to grow substantially.

[8] Trend Force (2024, February 28). [News] NVIDIA's H100 AI Chip no Lonher out of Reach, Inventory Pressure Reportedly forces Customers to resell. https://www.trendforce.com/news/2024/02/28/news-nvidi-as-h100-ai-chip-no-longer-out-of-reach-inventory-pressure-reportedly-forces-customers-to-resell/.

[9] Robison, K. (2024, February 22). Customer demand for Nvidia chips is so far above supply that CEO Jensen Huang had to discuss how 'fairly' the company decides who can buy them. Fortune. https://fortune.com/2024/02/21/nvidia-earnings-ceo-jen-sen-huang-gpu-demand-supply-allocate-fairly/.

[10] SCloud (2024, January 17). The Complex Case of GPU Smuggling: What, Who and How? https://www.scloud.sg/resource/the-com-plex-case-of-gpu-smuggling-what-who-and-how/.

GPU Supply-Demand Gap: AI Industry Challenges

The GPU industry is highly gatekept, with only a few enterprises controlling the overall available supply, making it an oligopoly. This small number of companies has significant market impact. Currently, there are three main producers of GPUs: Intel, AMD, and Nvidia. Nvidia and AMD both produce high-end GPUs that are predominantly in demand for database and AI applications, while Intel is more focused on the consumer market and is relatively new to the GPU scene. Consequently, Intel's GPUs are generally less optimized for AI computing and less in demand compared to Nvidia and AMD's offerings. Additionally, Intel's GPUs are generally weaker in terms of computing power.

Nvidia currently holds a near-monopoly on the GPU market with an 80% market share, making it a significant industry mover.[11] The recent shortage of supply has been primarily driven by a spike in demand from the AI sector, which has seen significant breakthroughs with LLMs leading to consumer-ready products like ChatGPT. This surge in demand has led to a gold rush mentality, with cloud providers, tech conglomerates, and AI startups frantically buying GPUs. This situation is exacerbated by pre-existing chip shortages caused by the COVID-19 pandemic, the Suez Canal blockage, and the ongoing trade war between the US and China, all of which have disrupted global supply chains.

According to TSMC, a major Taiwanese silicon producer, the shortage of Nvidia AI GPUs is expected to persist for the next 1.5 years. TSMC cites the shortage of chip-on-wafer-on-substrate (CoWoS) packaging capacity as the main reason for the continued scarcity.[12] The limited access to GPUs has directly impacted the global GPU distribution, with Nvidia being accused of unfairly prioritizing high-end cards for cloud providers, leaving AI startups at a disadvantage. Nvidia is currently under regulatory scrutiny, with multiple jurisdictions, including the European Union, the United States, and China, investigating the company for potential antitrust violations. The main complaints involve Nvidia's allegedly anticompetitive practices in the GPU and cloud services provider markets, particularly regarding the fair distribution of GPUs.[13]

For AI startups, accessing high-end GPUs can be extremely challenging due to their scarcity and the limited leasing opportunities. Amazon Web Services (AWS) allows consumers to rent high-end GPUs like the Nvidia A100, but at a significant cost of approximately $4.10 per hour. This expense is prohibitive for most AI startups, as they often require multiple GPUs depending on the model used. To put this in perspective, training a 175-billion-parameter model (such as GPT-3) requires over a terabyte of data to be kept in memory, exceeding the capacity of any single GPU and necessitating the splitting of the model across multiple GPUs. Additionally, computational complexity adds considerable time to processing requests. Running a single GPT-3 inference operation without exploiting parallel architecture would take at least 32 hours, rendering the model impractical. In practice, AI models run on GPUs with a large number of tensor cores, reducing inference time significantly. For instance, the Nvidia A100 can reduce GPT-3

[11] Nguyen, J. (2024, March 8). What do you need to know about Nvidia and the AI chip arms race. https://www.marketplace.org/2024/03/08/what-you-need-to-know-about-nvidia-and-the-ai-chip-arms-race/#:~:text=Nvidia%20has%2080%25%20control%20over,up%20265%25%20since%20last%20year.

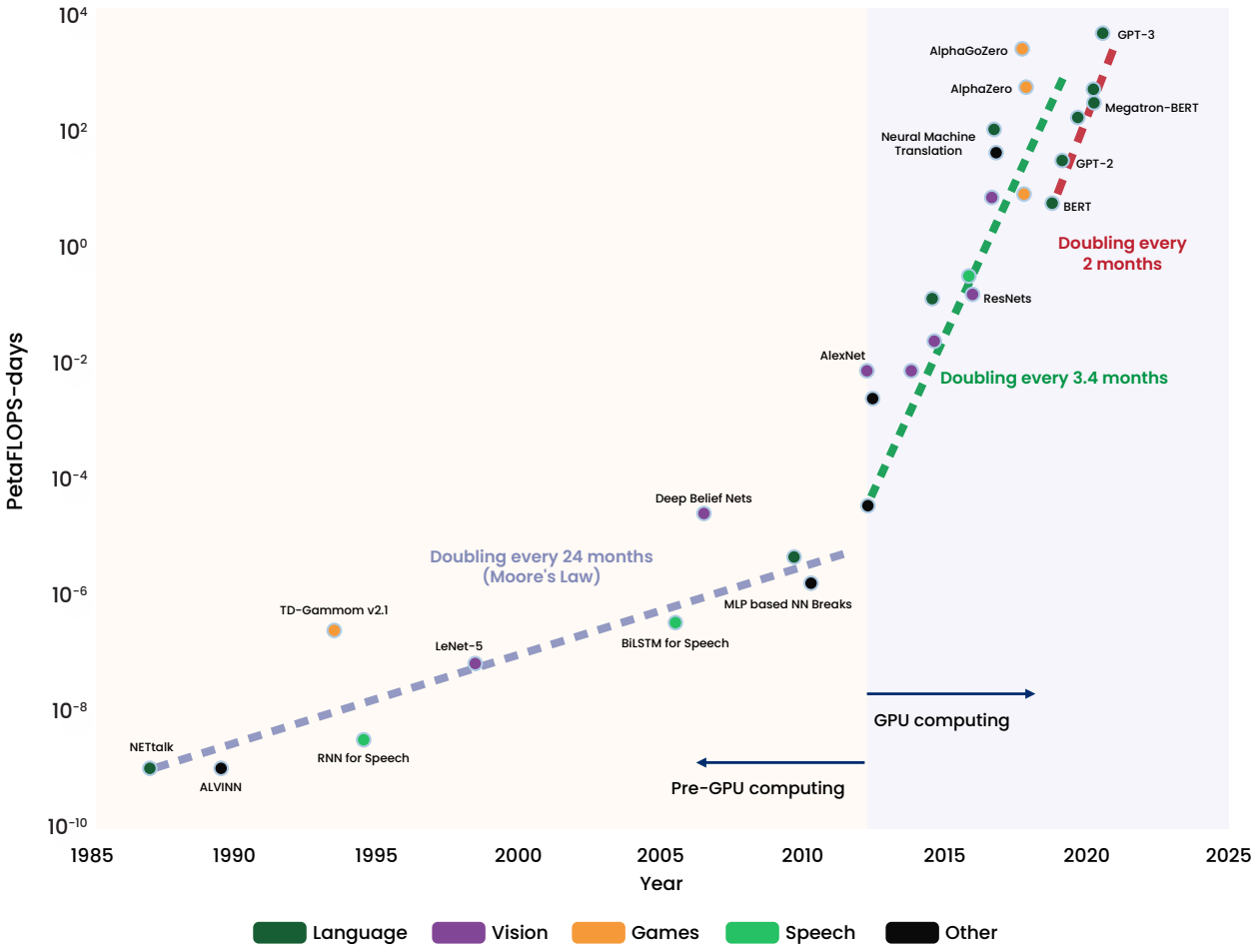
[12] Shilov, A. (2023, September 7). TSMC: Shortage of Nvidia's AI GPUs to Persist for 1.5 Years. Tom's Hardware. https://www.tomshardware.com/news/tsmc-shortage-of-nvidias-ai-gpus-to-persist-for-15-years.

[13] Reuters (2023, September 30). EU examines Nvidia-dominated AI chip market's alleged abuses, Bloomberg reports. https://www.reuters.com/technology/eu-starts-early-stage-probe-into-nvidia-dominated-ai-chip-market-abuses-2023-09-29/.

inference time to about 1 second. However, this is a simplified and generalized calculation, as the bottleneck is often not the GPU's computing power but the ability to retrieve data from the specialized graphics memory to the tensor cores.

The overall computational complexity and costs are substantial, with some AI companies reportedly spending more than 80% of their raised capital on computing resources.[14] The persistent gap between GPU supply and demand, driven by high costs, limited availability, and increasing computational requirements, continues to challenge the AI industry, particularly for startups striving to compete in this rapidly evolving field.

Figure 1 Computing Power Demand Overview over Time



Until 2012, computing power demand doubled every 24 months; recently this has shortened to approximately every two months

[14] Andreessen Horowitz (2023, April 27). Navigating the High Cost of AI Compute. https://a16z.com/navigating-the-high-cost-of-ai-compute/.

Centralized vs Decentralized Computing: Challenges and Constraints

Scaling a centralized database presents numerous challenges, particularly when access to Database Grade GPUs is limited. One major issue is the strain on network resources. As data centers expand, they demand an increasingly vast network infrastructure, which not only adds significant overhead but also heightens the likelihood of failures. This complexity necessitates robust, often redundant backup mechanisms, which in turn lead to substantial cost increases.[15]

Furthermore, the operation of large data centers requires a stable and high-voltage power supply, which is more complicated to manage than it might initially seem. Controlling such high voltages demands significant capital investment. Additionally, every watt of power used contributes to heat generation, necessitating advanced cooling solutions to manage the thermal output. These cooling processes, often reliant on extensive water usage, are expected to increase as data center operations expand. Financial considerations also impose limits on the scalability of data centers. Despite the technological capability to expand, there comes a point where it is no longer economically viable to do so due to escalating operational costs. For instance, in the Netherlands, the annual electricity consumption by data centers has surged dramatically, from 1,652 GWh in 2017 to an estimated 4,500 GWh in 2023. This amount of electricity could power approximately 1.8 million households for a year.[16] As the demand for resources like water for cooling continues to climb, the environmental impact and operational costs of data centers will inevitably rise, further challenging the scalability of centralized databases.

Decentralized data centers present a unique set of challenges, especially when compared to the centralized model of training LLMs on high-end GPUs in traditional data centers. A notable issue in decentralized setups is that they typically rely on consumer-grade CPUs and GPUs, which are contributed by the users themselves. This reliance significantly impacts the overall system performance due to the limited memory capacity of these devices. This constraint necessitates a more fine-grained partitioning of deep neural networks (DNNs), along with the distributed storage of datasets and the development of more efficient scheduling algorithms to optimize resource usage. The differences between consumer and database grade GPUs will be further highlighted in the next chapter.

Additionally, the wide variety of software and hardware configurations across geographically distributed devices complicates programming efforts. This diversity often leads to compatibility challenges that must be addressed to maintain a stable network. Moreover, the dynamic nature of decentralized networks, where devices can join or leave unexpectedly, introduces substantial challenges in terms of fault tolerance and requires robust, dynamic rescheduling mechanisms. Communication delays pose another significant challenge; low network bandwidth can result in unacceptably long communication times, which is particularly problematic given the substantial volumes of data exchanged between devices in decentralized systems.

Finally, the variability in hardware

[15] Butler, G. (2021, December 14). The challenges that arise with scaling up. Data Center Dynamics. https://www.datacenterdynamics.com/en/marketwatch/the-challenges-that-arise-with-scaling-up/

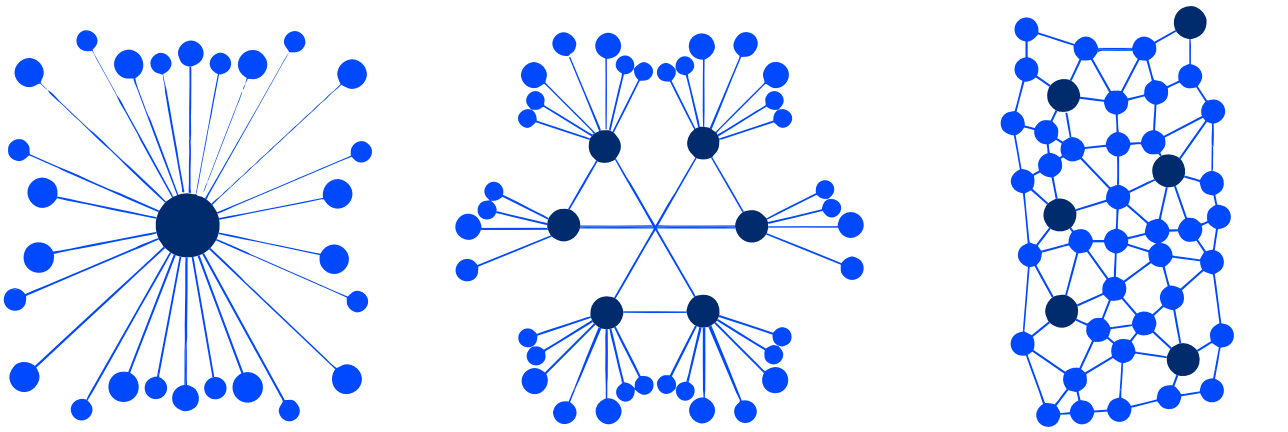
[16] Centraal Bureau voor de Statistiek (2022, December 8). Elektriciteit geleverd aan datacenters, 2017-2021. https://www.cbs.nl/nl-nl/maatwerk/2022/49/elektriciteit-geleverd-aan-datacenters-2017-2021

performance, marked by different GPU and CPU architectures, memory sizes, bandwidth, and battery capacities, further complicates the situation. This variability requires the consideration of diverse constraints to ensure that decentralized data centers operate efficiently and effectively. These multifaceted challenges highlight the need for innovative solutions to fully exploit the potential of decentralized computing frameworks.

Data Center Power Surge by 2026 from AI, Crypto

Data centers are projected to double their electricity usage by 2026, driven by AI and cryptocurrency mining, which demand immense power. In 2022, data centers consumed 460TWh, and this figure could soar to over 1,000TWh by 2026, equating to the entire power consumption of countries like Sweden or Germany. In the US, data center energy use is expected to rise from 200TWh to 260TWh, representing 6% of national power consumption. Ireland faces an even starker scenario, with data centers potentially consuming 32% of its electricity by 2026. Efforts to mitigate this surge include more efficient cooling technologies and legislative measures, but the challenge remains immense.

Figure 2 Overview of Centralized, Decentralized, and Distributed Networks.



Centralized Network
All the nodes are connected under a single authority

Decentralized Network
No single authority server controls the nodes, they all have individual entity

Distributed Network
Every node is independent and interconnected with each other

The Advantages and Limitations of Datacenter Grade- vs. Consumer GPUs

Datacenter Grade GPUs are powerful and high-end GPUs that deliver a high amount of computing power (i.e. FLOPS) in comparison with Consumer cards. Datacenter Grade GPUs often have a high amount of memory, and generally distinguish themselves of Consumer Grade cards in terms of type of memory. Database Cards often utilize ECC memory, which is considerably faster as GGDR or HBM memory. To put this in perspective, a RTX 4090, a high-end consumer card, can generate on average 82.58 TFLOPS (FP 32 Tensor Core) with 24 Gigabytes (GB) memory, while an A100, a high-end Datacenter Grade card, can generate 155.92 TFLOPS (FP 32 Tensor Core) with 80 GB. FP 32 Tensor Cores are specialized processors found in advanced GPUs, specifically designed to handle complex calculations involved in AI and deep learning tasks efficiently. Thus making these Datacenter cards significantly faster, and potentially more suitable for heavy computing operations, such as machine learning. Additionally, Datacenter Grade GPUs are often easier to cool as they utilize passive fans, which allows one to stack them side by side, thus allowing you to place more GPUs in a relatively small space.

As highlighted earlier in this paper, the main limitation of these high-end computing Datacenter cards is the overall limited availability, making them relatively hard to acquire in comparison with high-end consumer cards. Despite the differences in computing power and memory, another parameter that has to be considered is the overall price of these cards. The prices of these Database Grade GPUs are significantly higher as these of Consumer Grade Cards, if we utilize the same respective cards as earlier, the overall cost difference is up to 6 times. The reason for the price difference is caused by market segment differentiation, as datacenter grade cards are targeted to considerable cloud companies, whereby the overall disposable income is significantly higher. As, two RTX 4090 GPUs offer twice the TFLOPS Tensor Core of an A100 while costing only about 35% of its purchase price.

However, it's important to note that cloud companies are prohibited from using consumer GPUs for database operations due to the terms of service of Nvidia's GeForce.[17] This restriction means that the drivers provided by Nvidia for these GPUs cannot be used. GPU drivers are crucial software that act as intermediaries between the computer's operating system and the GPU, enabling the computer to utilize the GPU effectively. Without the appropriate drivers, the GPU may not function correctly, potentially affecting the performance of applications like games, videos, and graphic-intensive programs. Proper drivers ensure that the GPU and computer operate together seamlessly. Consequently, without them, GPUs may not achieve the expected FLOPs. This issue will be discussed in more detail in the risk assessment section.

The overhead costs associated with consumer GPUs are notably higher when compared to datacenter GPUs. To achieve the performance of a single datacenter card, multiple consumer cards are required, which significantly increases power consumption and, consequently, electricity expenses. This greater power usage also generates more heat, necessitating more robust cooling solutions that further elevate electricity costs. Consumer GPUs typically employ active cooling systems that use fans to expel heat. In contrast, many datacenter GPUs rely on passive cooling systems integrated within the server room's infrastructure, which are generally more efficient. This difference in

[17] Nvidia (n.d.). License for Customer use of Nvidia GeForce Software. https://www.nvidia.com/en-gb/drivers/geforce-license/.

cooling generally more efficient. This difference in cooling efficiency contributes to the higher operational costs of consumer cards. Additionally, the larger size of consumer GPUs restricts the number that can be installed in close proximity, complicating their deployment in the dense configurations typical of data centers.

Figure 3 Comparative overview of Consumer- and Database Grade GPUs and Model Memory

GPU	TFLOPS (FP32)	TFLOPS FP32 TENSOR CORE	MEMORY	LEVEL
RTX 4090	82.58	82.58	24GB	CONSUMER
RTX 4080	48.74	97.5	16GB	CONSUMER
RTX 3080	29.77	59.5	10GB	CONSUMER
H100	51.22	756	80GB	DATA CENTER
A100	19.49	155.92	80GB	DATA CENTER

Chapter Summary

In the first half of 2023, OpenAI's ChatGPT gained significant attention, underscoring the AI industry's rapid growth, currently valued at $180 billion and projected to exceed $800 billion by 2030. ChatGPT and other Large Language Models (LLMs) can generate and comprehend human language, enabling various applications. This progress has spurred major tech companies like Google and Meta to develop LLM-based solutions. Deep Learning (DL) algorithms for these models require substantial computational power, increasing the demand for high-end GPUs and CPUs.

Training LLMs involves billions of parameters, necessitating efficient data processing. This demand is prominent among tech giants, AI startups, and cloud providers, all vying for GPUs. Companies like Meta and Tesla have invested heavily in custom AI models and research, while Anthropic, Snowflake, and Databricks have acquired numerous GPUs to support AI development. To meet this demand, GPU providers like Nvidia have introduced hardware improvements, including integrated tensor cores for DL workloads. New AI Application-Specific Integrated Circuits (ASICs) such as Groq's Tensor Streaming Processors and Graphcore's Intelligence Processing Unit promise enhanced AI computing performance.

The surge in GPU demand has led to a supply crunch, with delivery times peaking at 11 months and stabilizing around 4 months. This has raised concerns about the fair allocation of GPUs, particularly accusations that Nvidia prioritizes cloud providers, disadvantaging AI startups. Regulatory scrutiny is increasing, with investigations into Nvidia's distribution practices in multiple jurisdictions. The GPU supply-demand gap remains a significant challenge, with the industry dominated by a few companies, creating an oligopoly. Nvidia, holding an 80% market share, has been particularly impacted by the AI-driven demand spike. This situation is exacerbated by pre-existing chip shortages from the COVID-19 pandemic, geopolitical tensions, and trade restrictions, leading to continued GPU scarcity. AI startups struggle to afford high-end GPUs like the Nvidia A100, costing about $4.10 per hour on Amazon Web Services (AWS).

Io.net offers a decentralized computing network, enabling ML and DL engineers to access distributed cloud clusters at lower costs than centralized services. This network provides high-end GPUs and CPUs, often inaccessible due to high costs and limited availability. Io.net aims to bridge the supply-demand gap, offering a more affordable and accessible alternative for AI startups and mid-sized companies. However, decentralized networks like io.net face challenges, including reliance on consumer-grade hardware, which impacts performance due to limited memory and diverse configurations. Despite these issues, decentralized networks offer a promising alternative to traditional centralized databases, addressing scalability and cost-efficiency concerns.

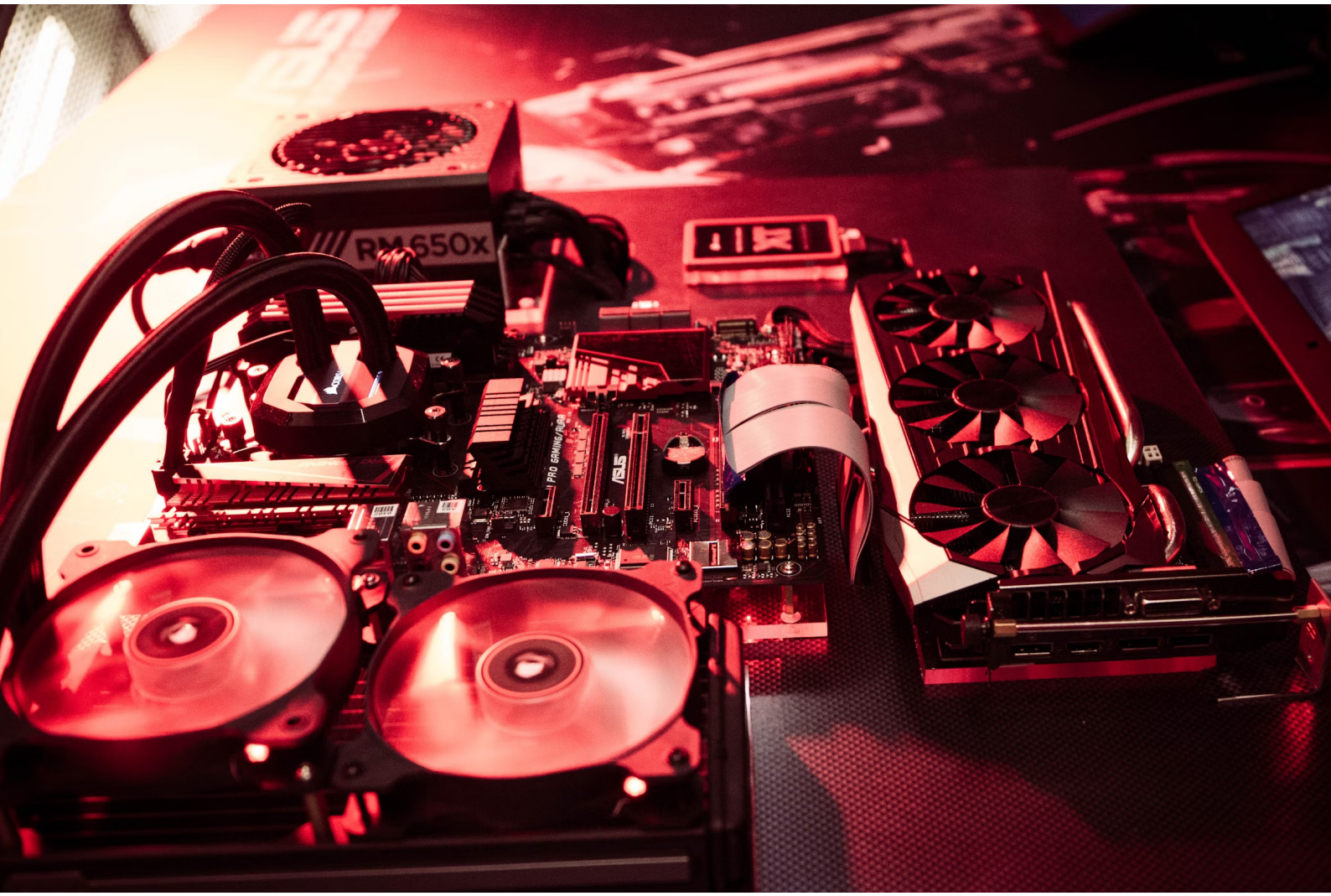


Photo by Maxime Rossignol on Unsplash

Introduction to io.net: Democratizing Distributed GPU Computing

Io.net provides affordable access to high-end GPUs for mid-sized companies and startups, leveraging a scalable, decentralized network to enhance computing capabilities.

Io.net is dedicated to democratizing access to high-end GPUs, including both consumer and database cards, for mid-sized companies and startups through the network of GPUs and CPUs connected to io.net. By offering these typically hard-to-acquire resources, io.net not only enables these organizations to enhance their computing capabilities but also aims to offer these services at prices significantly lower than those of its competitors. Anyone can join the io.net network by connecting their GPU to the io.net ecosystem, enabling them to share their underutilized computing resources with others. The process of joining the network will be detailed in the upcoming chapters.

Leveraging a scalable cluster architecture similar to that of Amazon Web Services (AWS), io.net organizes its resources into geographically distinct zones. Within these regions, availability zones (AZs) consist of data centers connected through a highly integrated network backbone. This setup allows io.net to distribute computing resources across multiple data centers within a region, mimicking the distributed nature of clusters provided by AWS. Io.net boosts its service reliability using two key mechanisms: autoscaling and fault tolerance.

Autoscaling is an essential feature in cloud computing that allows the system to automatically scale its computational resources based on current demand. This dynamic adjustment ensures the system can efficiently handle workload fluctuations without manual intervention, optimizing resource usage and reducing costs. Fault tolerance, on the other hand, ensures the system's continuous operation despite the failure of any components. Utilizing Kubernetes (K8s) as an example, the system demonstrates this capability effectively. If a node fails, K8s promptly requests a replacement from the Cloud Service Provider (CSP). It then swiftly deploys the required workload specifications to the new worker, ensuring minimal disruption and maintaining system stability and availability. Additionally, io.net employs "shadow workers" (i.e. standby GPU nodes) to further enhance system reliability and fault tolerance. Unlike active resources, shadow workers remain in standby mode, ready to take over immediately should any primary resource fail.

This setup ensures they are fully configured and can be activated without delay, providing a robust safety net for maintaining uninterrupted service. Moreover, to safeguard network integrity and ensure trust, io.net mandates that service providers commit io.net tokens as collateral. If a provider engages in malicious activities, they are subject to automatic penalties, including the programmatic slashing of their tokens. Specifically, if a node becomes unavailable or fails while in operation, the provider will forfeit an amount equivalent to one hour's worth of rewards. The financial penalty is relatively mild, which could create an imbalance since the consumer might experience significant damage. Therefore, integrating a reputational system with long-term effects could lead to a more balanced ecosystem. This concept will be further explored in the final chapter of this research, "Risks and Recommendations."

This policy aims to enhance reliability but also ensures the fair use of the network's resources. io.net addresses the issue of information asymmetry through its io.net Explorer tool, which aims to enhance transparency in the network. The explorer offers consumers comprehensive information about the available nodes. Since much of the platform's activity is conducted on-chain, transactions and operational statuses are visible and verifiable by all parties involved. This transparency is intended to enable consumers to access real-time data on the performance, reliability,

and cost-effectiveness of available GPU resources. The goal is to facilitate informed decision-making and foster a more equitable environment within the ecosystem.

A distinctive feature of io.net, compared with its Web3 counterparts, is the allowance of fiat payments. Therefore, allowing users to pay with fiat on-ramps significantly increases the platform's accessibility for those unfamiliar with cryptocurrencies. Additionally, in the background, the native IO token will be purchased, ensuring the overall marginal increase in demand for the IO token is maintained. While a fiat on-ramp offers clear advantages, it also presents risks, which will be discussed in the risk chapter.

Ensuring Privacy and Security in io.net's Distributed Computing Network

A critical aspect of io.net's service is the protection of consumer privacy, particularly important when using decentralized GPUs to run potentially valuable AI algorithms. To address this, io.net implements end-to-end encryption and ensures that the containers deployed on nodes are isolated from the file system, preventing unauthorized access to sensitive data. Containers encapsulate an application with all of its dependencies (like libraries and other binaries) into a single package that is isolated from other containers and the underlying operating system, enhancing security and portability.

For added security, io.net identifies and utilizes specific GPU cards that support Trusted Execution Environments (TEEs). TEEs provide a secure area within a computer's processor where data can be processed in isolation, safeguarding the operations from external interference, including the operating system itself. This ensures that data processed within the GPU remains confidential and secure.

Furthermore, io.net is committed to achieving Service Organization Control Type 2 (SCO2) compliance across its network. SOC2 is a rigorous set of criteria designed by American Institute of Certified Public Accountants (AICPA) to ensure that a service provider manages data securely, upholding the privacy and interests of its clients. This compliance covers five critical areas: security, availability, processing integrity, confidentiality, and privacy. By meeting SOC2 standards, io.net demonstrates its dedication to robust controls that protect data and confirm the network's operational effectiveness.

"Io.net democratizes access to high-end GPUs for startups, enhancing computing capabilities with scalable architecture, robust security, and transparent, cost-effective resource management."

Supply and Demand Dynamics in io.net's GPU Network

The current network of io.net has provided nearly 800,000 GPUs that connected with the network in the last 30 days, with 40,000 verified through Proof of Work (PoW). The PoW process ensures that a GPU is genuine and not counterfeit, aligning with the demand for high-end consumer and database cards, as there is minimal demand for lower-tier cards. The number of verified GPUs is gradually increasing, especially following a Structured Query Language (SQL) injection attack in late April 2024, which will be further explained in the next chapter. The number of GPUs has grown significantly in recent months due to the ecosystem announcing an airdrop, leading to a surge in GPUs connecting to io.net. However, as with any airdrop, the overall effect will likely diminish once the incentive structure ends, as continuous airdrops would significantly increase asset inflation. The airdrop could serve as a step-up mechanism, allowing computing providers to familiarize themselves with the ecosystem and potentially remain if there is sufficient demand. However, this is only the case if there is sufficient demand, otherwise, providers are financially incentivized to move platforms.

Currently, the demand for io.net remains relatively low, with an overall utilization rate between 1-5%. For comparison, the Akash Network has a GPU utilization rate of around 25% (i.e. ~100 GPUs), but with a significantly lower total number of available GPUs—approximately 425 at the time of writing. Despite Akash's higher utilization rate, io.net has more users in an absolute sense, with over 10,000 GPUs constantly available on the ecosystem.

The low utilization rate of io.net can be attributed to several factors. As a startup, io.net lacks brand recognition, and the reputation of Web3 applications is generally low, with potential Web2 users often associating cryptocurrency projects with scams and fraud. Additionally, io.net's pricing is not yet the most competitive on the market. For example, Lambda, a direct competitor, offers slightly more appealing pricing for high-end database GPUs. This significantly impacts the overall cash flow generated by the ecosystem, but it is not unique to io.net. The cash flow generated by decentralized applications (dApps) in the crypto space is relatively low. Currently, the number of daily blockchain wallet users is around 11 million.[18] Despite there being over 16,000 identified dApps on dAppradar—and likely even more not listed there—the overall generated cash flow remains modest.[19]

As the number of GPU providers increases, incentivized by receiving tokens for contributing their GPUs to the ecosystem, it is expected that io.net's prices might decrease. However, this efficiency drive will only be effective if there is a corresponding increase in demand. Without increased demand, GPU providers may shift to other platforms offering higher earnings. This is particularly crucial given that staking rewards are anticipated to be relatively low. If prices do come down, overall demand is expected to rise in line with increased brand recognition. Lower prices could make the platform more attractive to a wider user base, driving adoption and usage. Additionally, as more users engage with the platform, network effects could further enhance its value, creating a positive feedback loop.

In turn, this could attract even more GPU providers and users, enhancing the ecosystem's overall stability and growth potential.

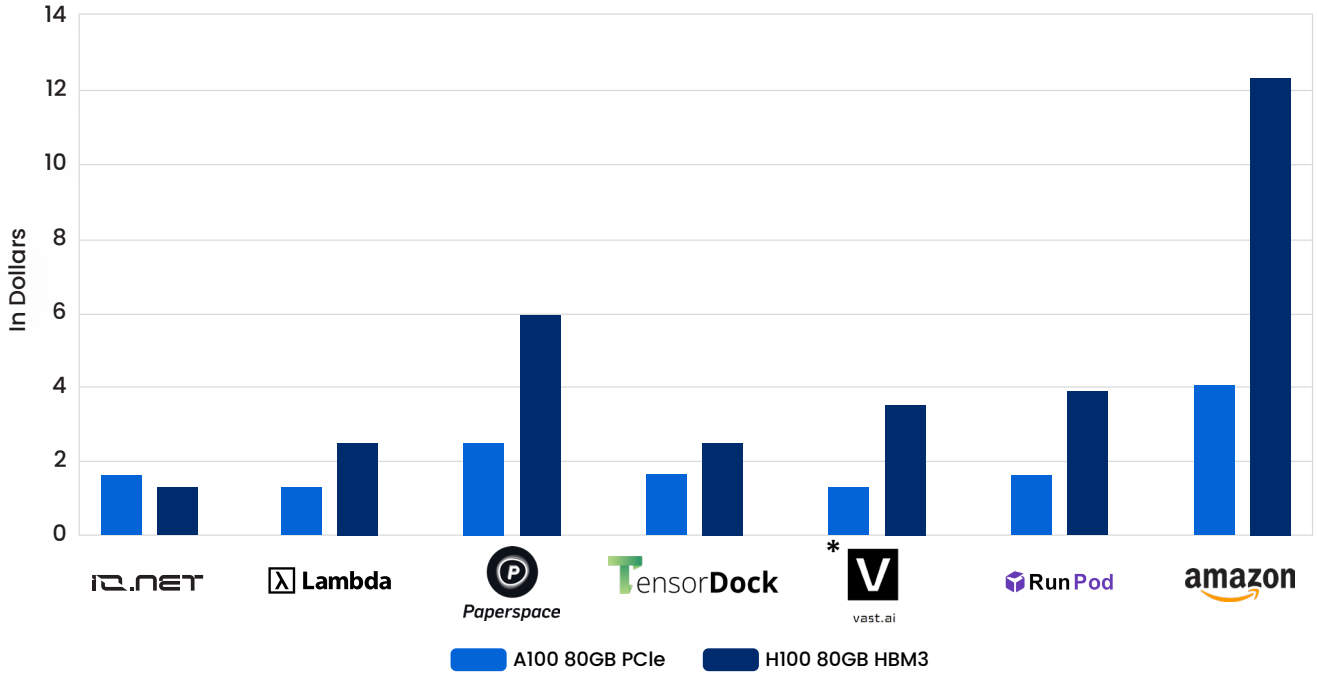
---

[18] Artemis (n.d.). Your No-Code Crypto Analytics Platform: Evaluate, compare and analyze trending chains and dApps across various metrics in one terminal. https://www.artemis.xyz/terminal.

[19] DappRadar (n.d.). Top Blockchain Dapps. https://dappradar.com/rankings.

Moreover, decentralized cloud providers can be less "plug and play" compared to traditional centralized providers, especially regarding library support and ease of integration. This lack of seamless integration can pose additional challenges for developers and users, potentially slowing down adoption rates. To address this io.net needs to focus on improving user experience by offering comprehensive documentation, robust developer tools, and extensive library support. Ensuring compatibility with popular frameworks and simplifying the onboarding process can make these platforms more accessible and attractive to a broader audience.

Figure 4 Comparison of Pricing for High-End GPUs Across Platforms



VastAi GPU is not as scalable as the other mentioned platforms.

Reliability and Stability of io.net's Ecosystem

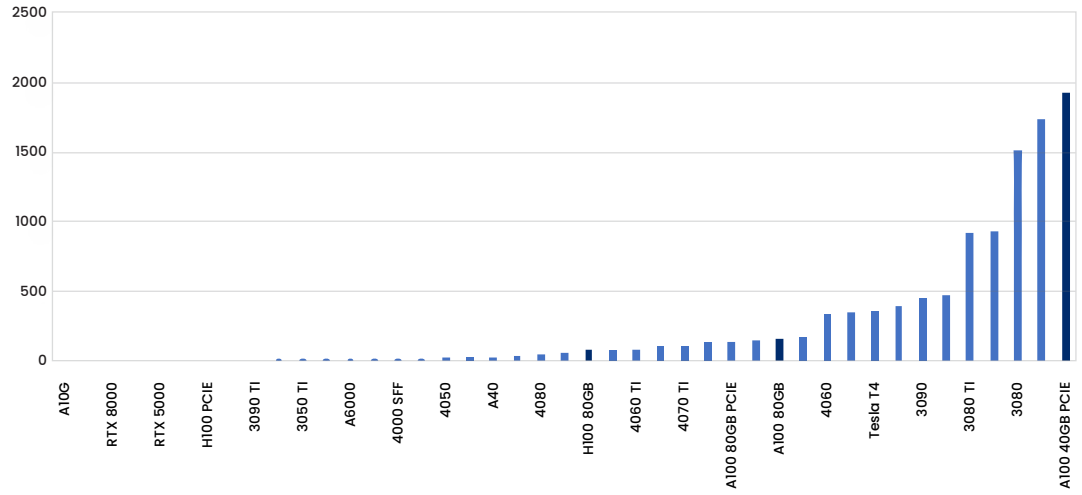
The io.net ecosystem has experienced its share of technical difficulties, with the most significant incident occurring on April 25th 2024. During this event, malicious actors exploited used ID tokens to perform a SQL injection attack, artificially inflating the number of GPUs. This prompted io.net to implement auto zero authentication (OKTA) at the device level. Auth0, a robust authentication service, was integrated to provide enhanced security features, ensuring the verification of both user and device identities before granting access to resources. With the integration of Auth0, each GPU in the network now undergoes rigorous authentication, ensuring that only verified and authorized devices can connect and operate within the ecosystem. This measure led to the necessity for all GPUs to authenticate, update, and restart, resulting in a significant temporary loss of GPU connections.

Despite the initial disruptions, the deployment of Auth0 authentication is expected to enhance the reliability of GPU providers by significantly reducing the risk of unauthorized access, thereby making the service more trustworthy. Auth0 also supports scalability, efficiently handling spikes in authentication requests without degrading performance, ensuring a smooth and reliable process under normal circumstances. However, it's important to note that Auth0 is a centralized system, with OKTA managing all authentication requests within the io.net ecosystem.

This centralization introduces a single point of failure for consumers and node operators relying on the authentication process. A notable incident occurred on May 3rd 2024, when a significant outage affected the io.net ecosystem as OKTA struggled to handle the volume of authentication requests, leading to cluster failures. Despite these challenges, the overall reliability of io.net remains significant under normal conditions.

House of Chimera has conducted multiple reliability and stress tests in April 2024, with different sets of GPU clusters. Apart from a few cluster issues, such as receiving 3060 TIs instead of 3080s, the clusters performed as expected. Each cluster utilizes an A4000 head node, which manages the underlying computational nodes. A common misconception is that head nodes are utilized for computational tasks; however, their primary function is management. The team has addressed the mislabeling issue where users received GPUs with similar but not identical performance characteristics. This mislabeling could impact memory types and sizes, affecting overall computational performance. Nonetheless, the process of setting up a cluster is straightforward, contributing to a relatively seamless user experience compared with traditional providers.

Figure 5 Cluster-Ready GPUs on IO.Net Platform



Snapshot is taken at 05/05/2024

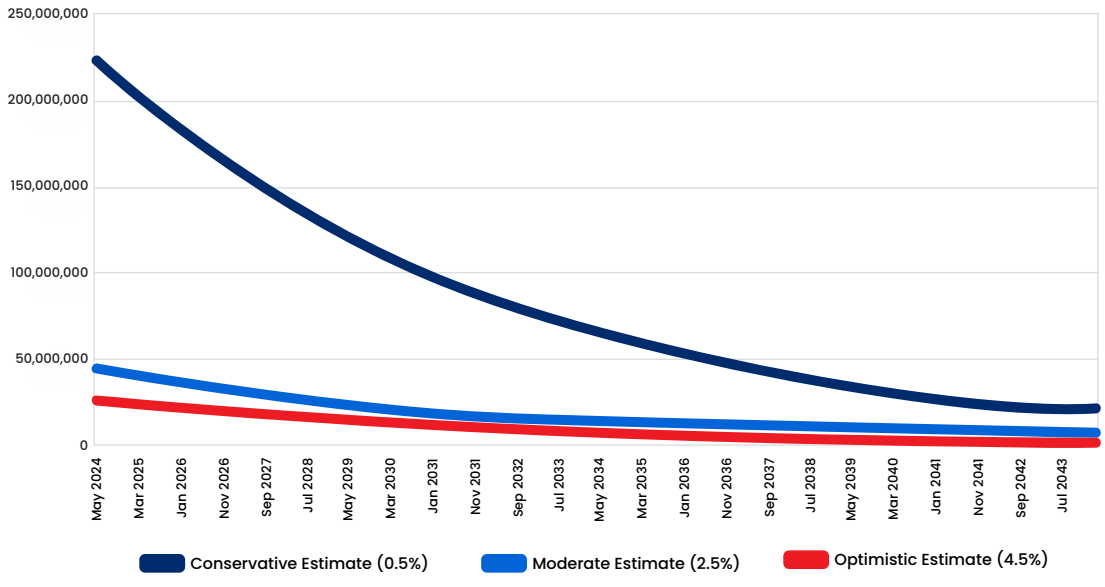
Reward Mechanisms in io.net's Decentralized Cloud Network

Io.net distinguishes itself from traditional cloud providers in several key ways, particularly through its decentralized architecture and its pre-payment requirement for services. Unlike conventional cloud services where payment is typically made after service usage, io.net requires users to pay upfront. This payment structure allows the autoscaler to effectively manage resources by tracking the available funds and allocating the appropriate computing resources accordingly. Pre-payment also secures fees upfront, which is crucial in a decentralized environment where recouping costs from a malicious actor post-service can be challenging. To further protect transactions, all fees are held in escrow until services are fully delivered, or they are refunded on a pro-rated basis if the consumer discontinues the service before the agreed date. This system ensures that both providers and consumers commit to their part of the service agreement.

Pricing within io.net is set by the ecosystem itself, meaning that there is a fixed price for every provided GPU. The rationale behind fixed pricing is that it is too early for the network to be completely decentralized and allow prices to be set by market demand, as the overall demand is still relatively low. However, the io.net team has ensured that in the future, GPU providers will be able to set their own prices based on supply and demand. This model supports a free market system without subsidies or other forms of market intervention, encouraging fair pricing that reflects the current value of the services offered.

Consumers can deploy clusters using the IO native token, fiat payments, or other supported network tokens. io.net has created structural demand for the IO token within the payments system. Computing providers can opt to receive payments in the native token or fiat. If the supplier prefers payment in the native token and the consumer pays differently, the consumer's funds will be used to acquire the native token, thereby creating demand. To incentivize payment in the native token, io.net charges transaction fees to both users and suppliers but does not charge any fees if the payment is made in the native token. Payments in fiat or other supported network tokens (i.e. non-native tokens) are subject to a 2% fee. Additionally, there is a reservation fee of 0.25% of the total cost to reserve the nodes. This fee is charged to both the supplier and consumer and cannot be waived even if payment is made in IO tokens.

Figure 6 Monthly Projected IO Tokens Required to Fully Offset Staking Rewards

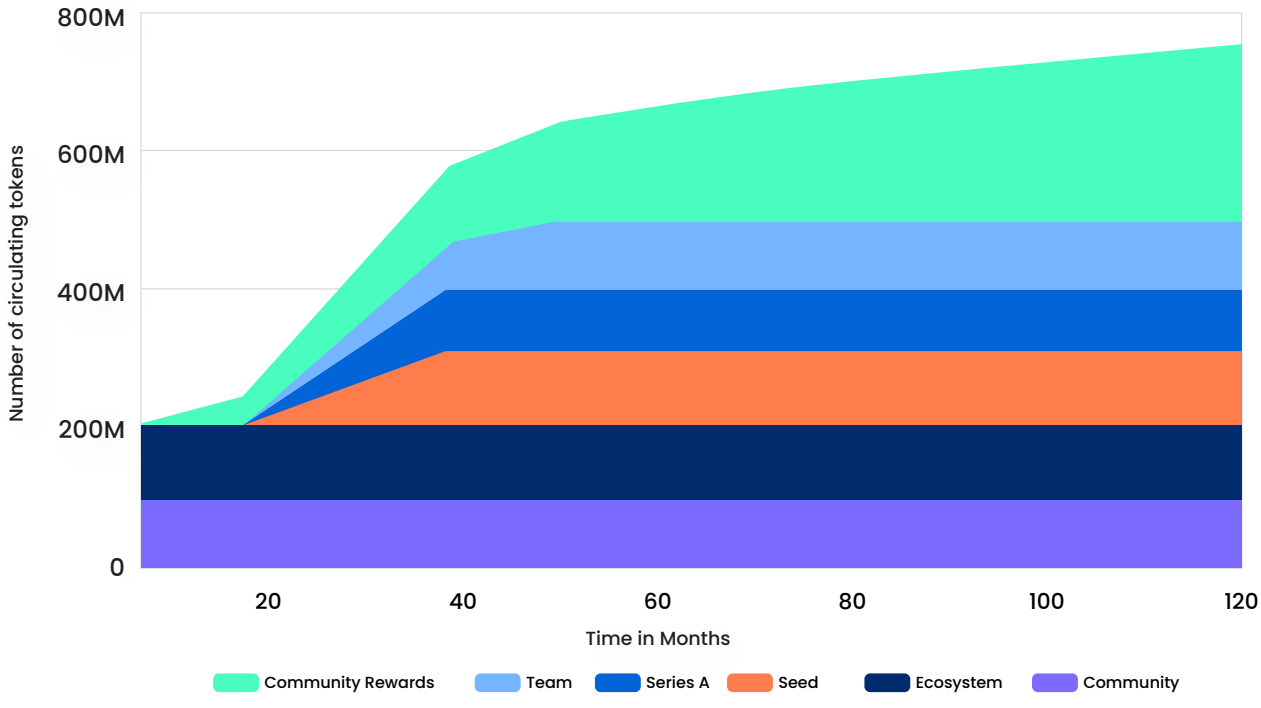


The Role of the IO Token

The IO Token is the utility token of the IO network, based on the Solana blockchain with a fixed maximum supply of 800 million tokens. Its primary utility is as a medium of exchange and to secure the underlying IO network. Although the IO token is not mandatory for transactions within the ecosystem, users can choose to use fiat currency to pay and receive payments, albeit with a 2% fee that is waived if the IO token is used. At genesis, the IO token will have an initial supply of 500,000,000 tokens, distributed across multiple baskets with different vesting schedules (see Appendix). The remaining 300,000,000 tokens will be allocated for community rewards. Specifically, these tokens will be emitted and paid to suppliers and their stakers as hourly rewards at a decreasing rate over time (see Appendix).

The rationale behind the decreasing rate is the expectation that GPU demand will increase over time. The staking rewards act as a subsidy, decreasing as the operational cash flow of the network increases, thus replacing the need for the subsidy. The emissions are designed to provide rewards to suppliers and stakers for 20 years, starting with an initial inflation rate of 8%, which decreases by approximately 12% each year. To partially offset the emission inflation, a fixed fee of 0.5% will be burned. Additionally, another 2% to 4% transaction fee could potentially be burned if the payment is made in fiat or other supported payment networks. To become a service provider, one must provide a minimum amount of IO token collateral, which must be staked for a node to receive IO idle rewards from the network. Users can add additional collateral by staking their tokens to a node, further increasing the network's security. In return for the perceived risk, users are compensated with staking rewards.

Figure 7 IO.net Token Vesting Schedule



Team Overview

The io.net team blends traditional and Web3 experience, with past roles at companies such as Facebook, Avalanche, Binance, and Capital One. They have held prominent positions like CMO and CTO, demonstrating their expertise. Presently, the team has approximately 50 members and plans to grow as operational cash flow improves. The C-level executives of io.net are shown in Figure 3 of the Appendix.

Ahmad Shadid, the founder, has been an entrepreneur since the age of 19. He got involved in crypto in early 2017 and became a Key Opinion Leader (KOL) in the emerging crypto community in the Middle East. His first venture, Arabfolio, was a crypto news platform and community investment syndicate. Ahmad has faced scrutiny for the outcome of Arabfolio, which experienced losses for Ahmad and investors due to crypto volatility. Allegedly, Shadid lost approximately 95% of Arabfolio's funds, as reported by multiple Arabic sources. House of Chimera used an AI translator to translate these sources, which may not be perfectly accurate, but is believed to be mostly correct. Despite thorough research, including reviewing public documents, House of Chimera found it challenging to verify these claims due to the language barrier and overall limited sources. However, multiple videos question Shadid's legitimacy, particularly concerning his previous venture, Arabfolio.

Shadid also has faced criticism for his second startup, WhalesTrader, a copy-trading platform that he co-founded with a partner. Due to disagreements with his co-founder at WhalesTrader, the co-founder launched a clone of the WhalesTrader business and Shadid decided to shut down the startup to focus elsewhere. We relied on an official statement from io.net regarding this case, as the WhalesTrader website is currently down and the web archive offers very limited information. We could not verify the claims made in the official statement through third-party sources.

Other ventures associated with Shadid include io.net, Antbit, and possibly Wborsa, all of which were fintech platforms except io.net. Antbit raised at least $2 million in funding and was rebranded as io.net, a pattern that might also apply to Wborsa, given the alignment of their fundraising dates with those of io.net.[20] The overall transition from a trading platform (e.g., "hedge fund in your pocket") to a distributed GPU platform remains unclear.

It is important to note that these claims were made over half a decade ago, so the overall validity of the claims remains unclear. What is clear is that Shadid held a managing role within Arabfolio, and multiple sources claim that Arabfolio experienced losses, though the extent of these losses is not well-defined. Additionally, the cryptocurrency industry is highly volatile, making losses a common occurrence. No OFAC Sanction list alerts were found on any of the Chief Managing staff, and no other negative press was discovered.

[20] Fintech Without Borders (n.d.). WBorsa – WorldBorsa Funding, Investor And Contact Details: San Francisco, California, United States. https://fintechwithoutborders.org/company/wborsa-worldborsa/

Chapter Summary

Io.net aims to democratize access to high-end GPUs for mid-sized companies and startups through a network of connected GPUs and CPUs. This network enhances computing capabilities at lower costs. Users can join by connecting their GPUs, sharing underutilized resources with others. The infrastructure, organized into geographically distinct zones similar to Amazon Web Services (AWS), uses autoscaling and fault tolerance mechanisms for reliability. Autoscaling adjusts resources based on demand, while Kubernetes (K8s) handles node failures by quickly replacing and deploying workloads. Shadow workers (standby GPU nodes) enhance reliability by taking over if primary resources fail.

Io.net requires service providers to commit io.net tokens as collateral to maintain network integrity and trust, with penalties for malicious activities, including token slashing. To address information asymmetry, io.net offers the io.net Explorer tool, providing real-time data on node performance, reliability, and cost-effectiveness, helping users make informed decisions. Unlike many Web3 platforms, io.net allows fiat payments, increasing accessibility for users unfamiliar with cryptocurrencies, though it introduces potential risks discussed later.

Ensuring privacy and security is a priority for io.net. End-to-end encryption and isolated containers prevent unauthorized access to sensitive data. Trusted Execution Environments (TEEs) offer additional security by processing data in isolation. Io.net aims for Service Organization Control Type 2 (SOC2) compliance, ensuring robust data management and security standards across its network.

The io.net network currently has nearly 800,000 GPUs connected, with 40,000 verified through Proof of Work (PoW) to ensure authenticity. Despite recent growth, the overall utilization rate remains low (1-5%), partly due to io.net's lack of brand recognition and competitive pricing compared to competitors like Lambda. As more GPU providers join, prices may decrease, potentially increasing demand. However, io.net needs to improve user experience and integration to attract a broader audience.

Io.net has faced technical challenges, including a significant SQL injection attack in April 2024, prompting the implementation of auto-zero authentication (OKTA) and enhanced security measures via Auth0. While these measures have improved reliability, they introduce a single point of failure due to centralization. Nonetheless, io.net continues to perform well under normal conditions, with House of Chimera conducting successful stress tests on different GPU clusters.

Io.net's founder, Ahmad Shadid, has faced scrutiny for past ventures like Arabfolio and WhalesTrader. Despite these controversies, no significant negative press or sanctions were found against the current managing team. Overall, io.net aims to provide a robust, decentralized GPU network, addressing the supply-demand gap in the AI and ML industries while ensuring security and reliability.

Photo by Growtika on Unsplash

Industry Analysis: The Competitive Landscape of Cloud Computing

Examining the growth, challenges, and opportunities in the cloud computing industry through the lens of the Five Forces Framework.

The cloud computing industry, encompassing both traditional and decentralized segments, has experienced significant growth in recent years. It is a dynamic and fiercely competitive landscape driven by the growing demand for high-performance computing (HPC), AI, ML, and data-intensive applications. At its core, this industry revolves around providing developers and organizations with access to scalable computing resources, including GPU computing power, delivered over the Internet. It is a dynamic and competitive landscape characterized by significant barriers to entry, intense competition among established players, and a strong position of hardware suppliers.

Traditional cloud computing giants like AWS, Google, and Azure dominate the market, leveraging their extensive infrastructure, brand recognition, and economies of scale, positioning themselves as trusted providers of high-performance computing solutions. This presents a major barriers to the remaining competition. Despite these barriers, innovative startups and niche players carve out opportunities by offering specialized services or pursuing decentralized approaches that promise affordability, flexibility, and more options for consumers (i.e., developers). In the decentralized cloud computing segment, platforms like Render Network, Akash Network, or Flux Network disrupt the traditional model by connecting consumers directly with GPU providers, bypassing centralized resource intermediaries. While decentralized providers face challenges in scaling their networks, being reliable, and accessing cutting-edge GPUs, they provide an alternative to developers seeking cost-effective computing resources.

In the following paragraphs, the industry is described in detail utilizing the Five Porter competition framework.

The Threat of New Entrants

CCloud computing, whether decentralized or centralized, is a difficult industry to penetrate considering the global tech giants involved in the space. First and foremost, an entrant must comply with data privacy laws and standards (e.g., SOC2, GDPR) given the developers may work with sensitive data to teach, for example, LLM models. This means one needs to understand, laws concerning consent, security and reporting requirements, cross-border transfers, and more.[21] Furthermore, providers must prioritize not only data but also infrastructure security. This has a severe impact on the brand, which is pivotal in building a network effect, where the service value increases as more customers and partners join. Additionally, the large players enjoy economies of scale making it hard for new entrants to compete on the scope of provided service and price.

Moreover, the established market participants have data center network across the globe requiring significant investment and stable cash flow. Consequently, centralized providers support popular GPU-accelerated frameworks and libraries for AI and ML, enabling seamless integration into developers' platforms, fostering innovation. To secure innovation, the firms must ensure significant funds dedicated to research and development expenditure and should make business partnerships with hardware producers (e.g., Nvidia) to deliver cutting-edge performance to customers. In addition, considering the decentralized computing providers, an entrant must differentiate from and compete with the traditional providers through price, a sufficient network of available GPUs requiring a fair and convenient reward mechanism for offering excess computing power.

Overall, the combination of high upfront investment, technological complexity,

[21] Eustice, J.C. (n.d.). Understand the intersection between data privacy laws and cloud computing. Thomson Reuters. https://legal.thomsonreuters.com/en/insights/articles/understanding-data-privacy-and-cloud-computing.

necessary infrastructure, regulatory requirements, network effects, and economies of scale creates significant entry barriers in the cloud computing market providing GPU computing power. However, innovative startups and niche players may still find opportunities by focusing on specific use cases, verticals, or regions where they can differentiate themselves and provide unique value to customers. Thus, an assumption can be made that the threat of new entrants is marginal.

Rivalry among Existing Competitors

Established providers with proven reliability and performance track records benefit from economies of scale, allowing them to spread fixed costs over a large customer base and offer competitive pricing. They can also negotiate favorable terms with hardware suppliers and pass on cost savings to customers. This environment can be described as an oligopoly characterized by high barriers of entry (e.g., infrastructure cost), product differentiation (i.e. scope of offered services), and interdependence where one firm's action impacts others. However, there is still room for innovation as the demand for HPC does not seem to vanish. An example of this can be Oracel Cloud, IBM Cloud, start-ups or even io.net which either differentiate by specialized services, industry expertise, focusing on niche verticals, or taking different approaches such as decentralization. Hence, an assumption can be made that the existing competition is severe, despite the developing cloud technology landscape, and that the threat is significant.

Power of Suppliers

Suppliers tend to have different scale of power depending on whether one focuses on centralized or decentralized computing power providers. Firstly, the traditional providers are severely dependent on GPU manufacturers considering the scarce resource the GPUs are nowadays. Furthermore, the tech giants are prohibited by the likes of Nvidia to utilize the consumer cards and must use the dedicated data center cards, which, on the contrary, are allegedly solely distributed to the tech giants, particularly by Nvidia. This suggests that both cloud computing and GPU manufacturing are intertwined industries.

However, the suppliers do not exclusively rely on the performance of the former since they generate sales from other segments as well. For instance, Nvidia reported that $15 billion of revenue was attributed to data centers; However, gaming, professional visualization and automotive were responsible for the remainder which accounted for $11.5 billion.[22] Considering that GPUs have no real substitutes, switching suppliers is costly, the number of suppliers, and the manufacturer's relative performance independence, the power of suppliers is major in the case of centralized cloud computing companies.

Secondly, turning to the decentralized providers, the platforms are profoundly dependent on the network of direct GPU providers connected considering they represent the backbone of computing resources for the customers. The GPU providers are not victims of robust lock-in mechanisms. They are mostly affected by the reward mechanisms established by projects (e.g., io.net, Akash Network, Render Network), thus allowing them to hold some bargaining power. Furthermore, it is important to note that the decentralized providers' computing power capability might be somewhat restrained by the availability of cutting-edge GPUs, especially the ones dedicated to data centers, which are both expensive and limited given the cards are plausibly directly allocated to tech giants.

[22] Nvidia (2024). 2024 NVIDIA Corporation Annual Review.
https://s201.q4cdn.com/141608511/files/doc_financials/2024/ar/NVIDIA-2024-Annual-Report.pdf

Power of Consumers

Cloud computing is intended for developers who need sufficient resources to fuel innovation. Performance, stability, breadth of services, support for frameworks and libraries, and the security of data and infrastructure, tied to the provider's reputation, are paramount to customers. Established players excel in branding, reliability, and offering a wide array of services, leveraging their strong market presence to retain customers. However, their solutions often come at a higher cost. Decentralized alternatives present viable options, especially for startups and researchers, offering resources at a more affordable price point. When opting for a decentralized provider, customers face various options, differing primarily in price and level of security measures. This suggests that while customers may lack significant bargaining power in the traditional market, they wield considerable influence when it comes to decentralized cloud computing providers.

The Threat of Substitutes

The availability of GPUs is essential for HPC; Thus, developers often do not have any other option than to mandate cloud computing service providers for the above-mentioned reasons. Another option to gain access to computing resources is through research centers such as universities although their capabilities are likely limited due to high service costs and capacity. Hence, an assumption can be made that the threat of substitutes is minimal to non-existent.

"In a fiercely competitive landscape, cloud computing giants dominate, but innovative startups and decentralized platforms carve out opportunities through affordability and flexibility."

Value Drivers in the GPU Cloud Computing Market

The demand for computing power in the cloud computing industry is on the rise, creating opportunities for decentralized providers like io.net to thrive alongside established players. This growth is fueled by various factors.

Developers, particularly those in regions with limited access to hardware, may access consumer GPUs at more competitive prices through decentralized solutions enabling developers to utilize the GPUs' cumulative power. This democratization of resources empowers developers to drive innovation across industries beyond just AI. Furthermore, as industries like IT, Construction or Manufacturing embrace rapidly evolving technologies (e.g., automatization, robotization), the demand for computing power becomes crucial.

However, this increased access to resources also comes with risks, as outlined in the Risks and Recommendations chapter. Additionally, the exponential growth of data generated from internet usage, social media, online transactions, and mobile devices further amplifies the need for cloud computing solutions. This is because of the increasing volume and variety of data the developers and other consumers must deal with. Thus, as centralized cloud computing might become less economical for numerous consumers, decentralized providers could profit from the escalating need for cost-efficient and potentially expanded computing capabilities.

Market Opportunities in the GPU Cloud Computing Market

The AI sector is poised for significant growth, driven by its popularity in recent fundraising activities.[23] This surge is primarily due to the potential for AI technologies to be marketed directly to consumers, tapping into a vast retail market. Present-day large LLMs like ChatGPT are sufficiently advanced to be integrated into everyday applications, enhancing their utility and adoption. As more people and businesses begin to rely on these AI solutions, the demand for computational resources is set to rise sharply.[24] This increase is particularly crucial as it pertains to globally distributed computational resources, reflecting the anticipated worldwide expansion in LLM usage, especially as costs decline, making these technologies more accessible in less developed regions. The adoption of LLMs is expected to boost productivity, especially in sectors that traditionally rely more on human labor than on capital, such as textiles, agriculture, construction, and mining. In such labor-intensive industries, as the marginal productivity of labor diminishes, the relative benefits of investing in capital become more apparent, promoting a shift towards more capital-intensive methods. This shift is indicative of the broader 'Artificial Intelligence Revolution' expected to predominantly impact labor-heavy industries.

Furthermore, according to neoclassical capital theory, countries typically have an abundance of either capital or labor.[25] Nations rich in capital tend to maximize its use until the marginal returns of capital fall below those of labor, prompting a shift towards greater labor utilization, and vice versa. Recent research supports this theory, suggesting that higher capital abundance leads to increased automation, potentially reducing the need for labor and suppressing wages in labor-intensive fields where technology could replace manual tasks.[26] This dynamic underscores the transformative potential of AI in reshaping economic

[23] Thormundsson, B. (2024, March 27). AI startup company funding worldwide 2020-2023, by quarter. Statista. https://www.statista.com/statistics/1344128/worldwide-artificial-intelligence-startup-company-funding-by-quarter/.

[24] Mehonic, A., & Kenyon, A. J. (2022). Brain-inspired computing needs a master plan. Nature, 604(7905), 255-260.

[25] ScienceDirect (n.d.). Neoclassical Theory. https://www.sciencedirect.com/topics/social-sciences/neoclassical-theory.

[26] Acemoglu, D. (2024). Capital and Wages. Massachusetts Institute of Technology, 1-28.

landscapes, particularly in sectors most susceptible to automation.

An example of AI's influence extending beyond traditional industries is evident in the entertainment sector, particularly through cloud gaming. The rapid growth of cloud gaming represents a burgeoning opportunity within this expanding industry. According to PwC, gaming is among the fastest-growing industries, and cloud gaming is projected to drive significant value within this sector by 2027.[27] Cloud gaming leverages advanced computational resources to offer users unrestricted choice in gaming options while providing developers enhanced flexibility in game design. For instance, Nvidia's cloud gaming service, GeForce Now, reported over 25 million registered users in 2023, underscoring the rising popularity of this platform.[28] Similarly, Statista indicates that 7% of U.S. gamers engaged with cloud gaming platforms like GeForce Now in 2023, highlighting its increasing adoption.

The first major attempt at a cloud gaming platform was Google Stadia by Alphabet. Despite its early promise, the platform struggled with high latency issues, particularly affecting games requiring quick reflexes such as First-Person Shooters (FPS) (e.g. Counterstrike) and Multiplayer Online Battle Arena Video Games (MOBAs) (e.g. League of Legends). The challenges were partly due to the high-speed internet and high-end computing resources that were less common in 2018. Google Stadia was discontinued in early 2023.[30] In the Web3 space, the Metaverse is a significant trend, with companies like Meta investing heavily—46 billion USD —in developing their virtual reality worlds where users can interact.

This development parallels the challenges in high-end gaming, where advanced graphics require powerful GPUs, which are not affordable for a large portion of the global population. This economic barrier limits the reach of developers and potentially cuts off significant revenue from millions of users worldwide.

By contrast, mobile gaming thrives on accessibility; a vast majority of the global population owns a smartphone capable of running games under a 'freemium' model, where basic play is free but advancement costs money. This model benefits from the broad availability of mobile phones across various price points, capturing a large market. The primary advantage of cloud gaming is that it eliminates the need for players to own expensive hardware like high-end GPUs. Instead, players can access these resources remotely, often through a subscription or pay-as-you-go model, effectively democratizing access to sophisticated gaming by reducing upfront costs.

This model not only makes gaming more accessible but also allows developers to create more complex and resource-intensive games and platforms. Examples include the Metaverse, which becomes more economically viable and broadens its market reach. By integrating AI and cloud gaming, we can observe a broader trend where advanced technologies enhance productivity and accessibility across various sectors, from labor-intensive industries to high-end entertainment. This synergy highlights the transformative potential of AI and cloud computing in shaping future economic landscapes.

---

[27] Wakelin, J., & Baker, A. (2024, January 16). Top 5 developments driving growth for video games. PWC. https://www.pwc.com/us/en/tech-effect/emerging-tech/emerging-technology-trends-in-the-gaming-industry.html#:~:text=The%20video%20game%20industry%20is,to%20%24312%20billion%20in%202027

[28] Andric, D. (2023, May 11). How many people use GeForce Now? - 2024 statistics. Levvvel. https://levvvel.com/geforce-now-statistics/.

[29] Clement, J. (2023, May 24). GeForce Now subscription rate among gamers in the U.S. 2023, by type. Statista. https://www.statista.com/statistics/1386263/us-geforce-now-gaming-subscription-rate-type/.

[30] https://stadia.google.com/gg/.

[31] Levy, A. (2024, April 1). Meta Platforms has spent $46 Billion on the Metaverse Ssnce 2021, but it's spending twice as much on this 1 Thing. The Motley Fool. https://www.fool.com/investing/2024/04/01/meta-platforms-has-spent-46-billion-on-the-metaver/#:~:text=Meta%20Platforms%20Has%20Spent%20%2446,1%20Thing%20%7C%20The%20Motley%20Fool.

Chapter Summary

The GPU cloud computing industry, both traditional and decentralized, has experienced substantial growth driven by the demand for high-performance computing (HPC), AI, ML, and data-intensive applications. This sector provides scalable computing resources, including GPU power, over the Internet. It is highly competitive, with significant barriers to entry, intense competition among established players, and strong influence from hardware suppliers. Traditional cloud computing giants like AWS, Google, and Azure dominate the market through extensive infrastructure, brand recognition, and economies of scale. They offer high-performance solutions that are hard for newcomers to match. Despite these barriers, innovative startups and niche players find opportunities by offering specialized services or decentralized approaches that promise affordability and flexibility. Decentralized platforms like Render Network, Akash Network, and Flux Network connect consumers directly with GPU providers, bypassing centralized intermediaries. While these platforms face scaling and reliability challenges, they provide cost-effective alternatives for developers.

Threat of New Entrants: The cloud computing market is difficult to penetrate due to the dominance of global tech giants, compliance with data privacy laws, infrastructure security, and high upfront investment requirements. However, innovative startups can find opportunities by focusing on specific use cases or regions.

Competitive Rivalry: Established providers benefit from economies of scale and competitive pricing, creating an oligopoly. Despite high entry barriers, there is room for innovation, as shown by specialized services from Oracle Cloud, IBM Cloud, and startups. The competition remains fierce.

Supplier Power: Traditional providers depend heavily on GPU manufacturers like Nvidia, which have substantial bargaining power due to the scarcity of GPUs. Decentralized providers rely on a network of direct GPU providers, who hold some bargaining power influenced by reward mechanisms.

Consumer Power: Developers need reliable, high-performance computing resources. While traditional providers excel in branding and reliability, decentralized alternatives offer more affordable options, giving consumers significant influence in the decentralized market.

Threat of Substitutes: The essential nature of GPUs for HPC means there are minimal substitutes. Research centers offer an alternative, but their capabilities are limited, making the threat of substitutes negligible.

Photo by Markus Spiske on Unsplash

Risks and Recommendations

Navigating regulatory challenges, pricing inefficiencies, and enhancing reliability to ensure the stability and growth of the io.net ecosystem.

One of the primary challenges facing the io.net ecosystem is the risk of regulatory intervention, particularly in scenarios where a government mandates the cessation of services to a specific consumer or node. This concern arises from the decentralized nature of the platform. For instance, if a consumer engages in illegal activities, it is unclear who is responsible for discontinuing their services. Presently, io.net's structure is not fully decentralized, which ironically mitigates this risk to some extent. As pointed out, the Auth0 process centralizes control, enabling io.net to revoke the Auth0 credentials of a node, effectively disconnecting it from the platform. This centralized control allows io.net to also remove nodes from its system through the user interface.

However, the io.net team is committed to achieving full decentralization, which presents significant challenges. Another potential regulatory issue involves the integration of a Web2 fiat on-ramp without corresponding Know Your Customer (KYC) checks for consumers. This could allow malicious users in regions with scarce GPU resources to anonymously hire node operators, potentially facilitating money laundering within the IO ecosystem. While this is a risk common to many decentralized networks, it could be mitigated by implementing Decentralized Identities (DIDs). DIDs would enable KYC verification while maintaining user privacy by requiring only necessary personal data to be shared with specific parties. A Web3 project exemplifying this approach is idOS, which offers identity services that allow decentralized applications (dApps) to integrate through its idOS SDK. This enables users to manage their data and maintain control over their personal information.[32]

Additionally, regulatory challenges could arise from the international embargo on high-end GPUs, such as the A100 and H100 models from Nvidia, to certain countries like China. Despite these bans, platforms like io.net could inadvertently facilitate access to these restricted technologies due to the lack of integrated KYC procedures. This situation is particularly precarious as it might lead to significant legal repercussions, especially given the U.S. government's stringent controls, including export restrictions to the Middle East aimed specifically at preventing such technologies from reaching China.[33,34] These complexities underscore the delicate balance io.net must maintain between advancing technological innovation and adhering to international regulatory standards.

[32] idOS Docs (2024, April). Welcome to the idOS: The Identity Layer of Web3. https://docs.idos.network/idos-docs.

[33] Baptista, E. (2024, January 15). China's Military and Government acquire Nvidia Chips despite US Ban. Reuters. https://www.reuters.com/technology/chinas-military-government-acquire-nvidia-chips-despite-us-ban-2024-01-14/.

[34] Shilov, A. (2023, August 30). U.S. Bans Sales of Nvidia's H100, A100 GPUs to Middle East. Tom's Hardware. https://www.tomshardware.com/news/us-bans-sales-of-nvidias-h100-a100-gpus-to-middle-east.

Drawbacks of a Fixed Pricing Model

The current pricing model on the io.net platform is dictated by the io.net team, creating a non-free market scenario where supply providers can't dynamically adjust their prices based on demand. This fixed pricing leads to market inefficiencies, as prices do not respond elastically to changes in demand or supplier costs. This inflexibility can make it less profitable for suppliers to offer their services, potentially reducing supply, or prevent suppliers from taking advantage of higher demand through increased prices. In an ideal free market, prices would naturally fluctuate based on the interplay of supply and demand, ensuring efficient resource allocation. Centralized pricing, however, can lead to either shortages if prices are too low or surplus if prices are too high, disrupting market balance.

Introduction of Reputational Slashing

The reliability of the io.net platform is critically important for any computing application, especially AI platforms that require high stability and performance. To enhance reliability, several recommendations can be made. First, implementing robust filter options that allow users to select nodes based on specific criteria, such as uptime and performance (e.g., only nodes with at least 95% uptime), would help users match their computing needs with the most reliable nodes available.

Second, developing a reputational scoring system for suppliers, assigning scores based on the historical performance and reliability of their nodes, in a trustless manner based on a preset algorithm, would provide users with an additional layer of assurance. Higher scores would indicate more trustworthy suppliers.

Third, introducing reputational slashing mechanisms to penalize suppliers who fail to meet performance obligations is essential. Reputational slashes would negatively impact supplier scores, encouraging high performance and reliability. Unlike financial penalties, which may not have long-term effects, a damaged reputational score has lasting consequences, potentially protecting consumers more effectively than financial punishments alone. Stability and reliability are paramount in the computing industry, particularly for AI applications that require consistent and dependable resources. By implementing these recommendations, io.net can efficiently address the stringent demands of AI and other high-performance computing applications, ensuring a stable and trustworthy environment for its users.

Chapter Summary

The io.net ecosystem faces significant regulatory challenges, primarily due to its decentralized nature. One major concern is the risk of regulatory intervention if a government mandates the cessation of services to a specific consumer or node. This issue arises from the platform's current partial decentralization, allowing io.net to revoke the Auth0 credentials of a node, thus disconnecting it from the platform. However, io.net aims for full decentralization, which complicates this control.

A potential regulatory issue involves the integration of a Web2 fiat on-ramp without corresponding Know Your Customer (KYC) checks. This could allow malicious users in regions with scarce GPU resources to anonymously hire node operators, potentially facilitating money laundering. This risk, common to many decentralized networks, can be mitigated by implementing Decentralized Identities (DIDs) that enable KYC verification while maintaining user privacy. An example of this approach is idOS, which allows decentralized applications to integrate identity services through its SDK, enabling users to manage their data and control their personal information.

Regulatory challenges also stem from international embargoes on high-end GPUs, such as Nvidia's A100 and H100 models, to certain countries like China. Despite these bans, platforms like io.net could inadvertently facilitate access to these restricted technologies without integrated KYC procedures. This poses significant legal risks, especially given U.S. export restrictions aimed at preventing these technologies from reaching certain regions. io.net must balance technological advancement with adherence to international regulatory standards

The current fixed pricing model on io.net is set by the io.net team, leading to market inefficiencies. This model prevents suppliers from adjusting prices based on demand, causing potential supply reductions or missed opportunities to capitalize on higher demand. An ideal free market would allow prices to fluctuate naturally, ensuring efficient resource allocation. Fixed pricing can lead to either shortages or surpluses, disrupting market balance.

Reliability is crucial for io.net, especially for AI applications requiring high stability and performance. To enhance reliability, io.net could implement robust filter options allowing users to select nodes based on criteria such as uptime and performance. Additionally, developing a reputational scoring system for suppliers, based on historical performance and reliability, would provide users with added assurance. Introducing reputational slashing mechanisms to penalize suppliers failing to meet performance obligations could further ensure high performance and reliability. Unlike financial penalties, reputational damage has lasting consequences, encouraging suppliers to maintain high standards. By adopting these recommendations, io.net can address the stringent demands of AI and other high-performance computing applications, ensuring a stable and trustworthy environment for users.

Appendix

Figure 1 Token Allocation Overview of $IO

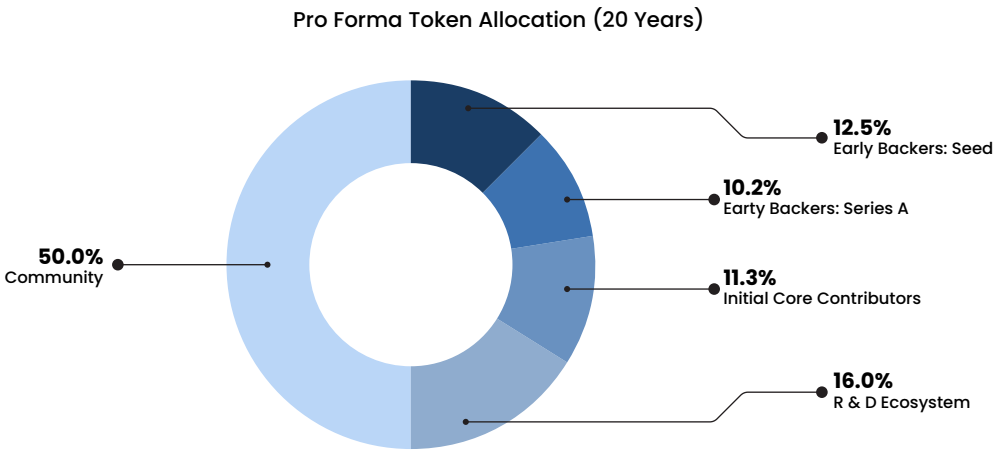


Figure 2 Yearly Inflation rate of IO.net based on the circulating supply

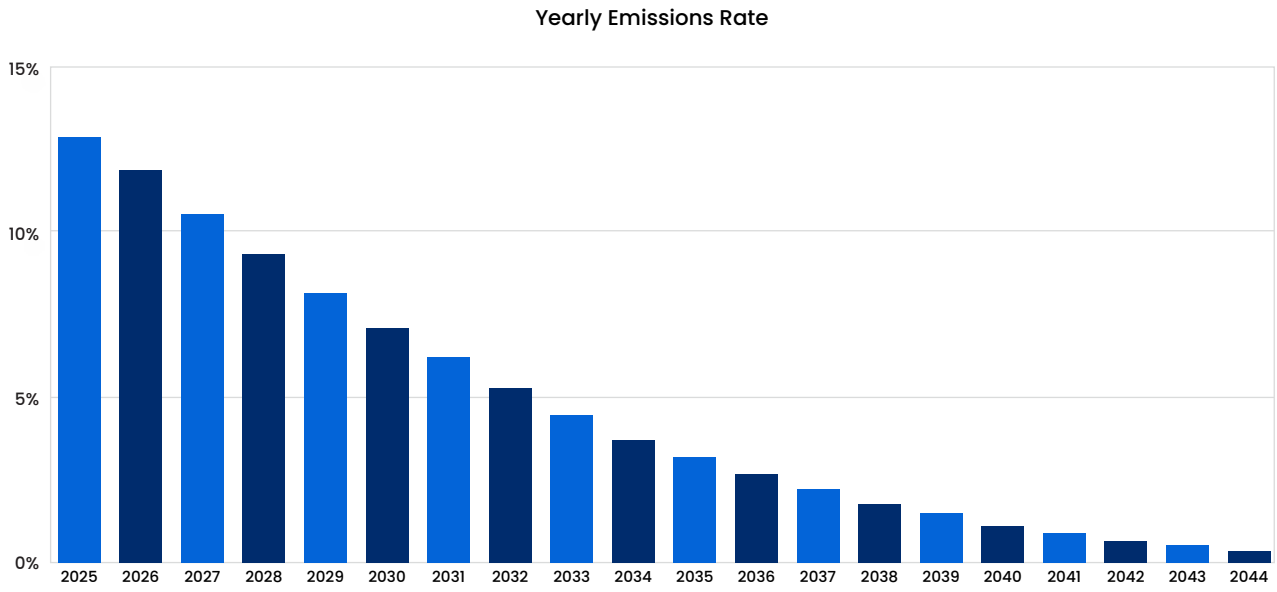
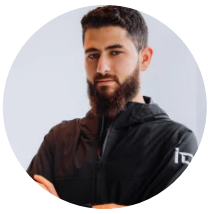


Figure 3 C-level Overview and Team Experiences of io.net



Ahmad Shadid

Tory Green

Basem Oubah

Gaurav Sharma

Hushky

Maher Jilani

Previous Experiences

BINANCE OAKTREE CapitalOne Ava Labs.

amazon kraken BAIN & COMPANY

crypto.com facebook

May 2024

Designed by House of Chimera

HouseOfChimera.com